Analysis of Diabetic Patient Hospital Records

UNC Charlotte

By: Husari Ghaith

Tanya Peddi

Table of Content

Abstract 4

**INTRODUCTION** 5

**DATA ANALYSIS** 5

**DATA CLEANING** 8

**Chi-square** 12

**CORRELATION** 12

**Linear Regression** 15

**PRINCIPAL COMPONENT ANALYSIS** 16

**Additional Prediction TASKS** 17

**Task 1: Bayesian Network** 17

**Task 2 – Hazard Function** 20

**Task 3: Text Analysis** 21

**CONCLUSION** 22

**References:** 22

R Code 23

Table of figures

[Figure 1:Mean of Numerical values 8](#_Toc482271250)

[Figure 2:Examples of data before cleaning 9](#_Toc482271251)

[Figure 3: Example of xtab() results 13](#_Toc482271252)

[Figure 4:Chi-Square result of Repaglinide against Readmission 14](#_Toc482271253)

[Figure 5: Chi-Square result of Insulin against Readmission 14](#_Toc482271254)

[Figure 6: Chi-Square result of Gender against Readmission 15](#_Toc482271255)

[Figure 7: Chi-Square result of Race against Readmission 15](#_Toc482271256)

[Figure 8: Chi-Square result comparison 16](#_Toc482271257)

[Figure 9: Linear Regression Result 17](#_Toc482271258)

[Figure 10: R excerpt of PCA numerical attributes 17](file:///C:\Users\ghusari\Downloads\New%20folder%202017%20downloads\lab11\Report.docx#_Toc482271259)

[Figure 11: R experpt of Index-values PCA analysis 18](file:///C:\Users\ghusari\Downloads\New%20folder%202017%20downloads\lab11\Report.docx#_Toc482271260)

[Figure 12: PCA analysis MIC and RMSR 18](file:///C:\Users\ghusari\Downloads\New%20folder%202017%20downloads\lab11\Report.docx#_Toc482271261)

[Figure 13:Bayesian Network of highly correlated variables 19](#_Toc482271262)

[Figure 14:Bayesian Network of all categorical variables 20](#_Toc482271263)

[Figure 15: Bayesian Network of all medicine related variables 21](#_Toc482271264)

[Figure 16: Bayesian Network 21](#_Toc482271265)

Abstract

We were provided with data of 10000 diabetic patients by a Health research organization and were asked to find any valuable information related to readmission of patients. We first analyzed the data and discovered there classes and levels. The data was filled with missing values, wrong variable classes and discrepancies which led us to performing data cleaning functions. To find the correlations between the data, we performed various tests and using the results we performed three major tasks. The first task involved finding the Bayesian Network of the system and thus finding what genuinely affects readmissions. This was followed by the second major task of calculating the hazard function and lastly the third task of text analysis of the diagnoses provided in the data.

**INTRODUCTION**

**DATA ANALYSIS**

The present analysis of a large clinical database was undertaken to examine historical patterns of diabetes care in patients with diabetes admitted to a US hospital and to inform future directions which might lead to improvements in patient safety. Databases of clinical data contain valuable but heterogeneous and difficult data in terms of missing values, incomplete or inconsistent records, and high dimensionality understood not only by number of features but also their complexity [1].

This study used the Health Facts database (Cerner Corporation, Kansas City, MO), a national data warehouse that collects comprehensive clinical records across hospitals throughout the United States. The database contains data systematically collected from participating institutions electronic medical records and includes encounter data (emergency, outpatient, and inpatient), provider specialty, demographics (age, sex, and race), diagnoses and in-hospital procedures documented by ICD-9-CM codes, laboratory data, pharmacy data, in-hospital mortality, and hospital characteristics [1].

It should be noted that the diabetic encounters are not all encounters of diabetic patients but rather only these encounters where diabetes was coded as an existing health condition. Since we are primarily interested in factors that lead to early readmission, we defined there admission attribute (outcome) as having two values: “readmitted,” if the patient was readmitted within 30 days of discharge or “otherwise,”which covers both readmission after 30 days and no readmission at all [1].

Nominal Data

* Race: Caucasian, Asian, African American, Hispanic, and other
* Gender: male, female, and unknown/invalid
* Age: Grouped in 10-year intervals: [0,10), [10,20), ...
* Admission type: Integer identifier corresponding to 9 distinct values, for example, emergency, urgent, elective, newborn, and not available
* Discharge disposition: Integer identifier corresponding to 29 distinct values, for example, discharged to home, expired, and not available
* Admission source: Integer identifier corresponding to 21 distinct values, for example, physician referral, emergency room, and transfer from a hospital
* Payer code: Integer identifier corresponding to 23 distinct values, for example, Blue Cross\Blue Shield, Medicare, and self-pay
* Medical specialty: Integer identifier of a specialty of the admitting physician, corresponding to 84 distinct values, for example, cardiology, internal medicine, family\general practice, and surgeon.
* Diagnosis 1: The primary diagnosis (coded as first three digits of ICD9); 848 distinct values
* Diagnosis 2: Secondary diagnosis (coded as first three digits of ICD9); 923 distinct values
* Diagnosis 3: Additional secondary diagnosis (coded as first three digits of ICD9); 954 distinct values
* Glucose serum test result: Indicates the range of the result or if the test was not taken. Values: “>200,” “>300,” “normal,” and “none” if not measured
* A1C test result: Indicates the range of the result or if the test was not taken. Values:“>8” if the result was greater than 8%, “>7” if the result was greater than 7% but less than 8%, “normal” if the result was less than 7%, and “none” if not measured.
* Change of medications: Indicates if there was a change in diabetic medications (either dosage or generic name). Values:“change” and “no change”
* Diabetes medications: Indicates if there was any diabetic medication prescribed. Values:“yes” and “no”
* 24 features for medications: For the generic names
  + Metformin
  + repaglinide,
  + nateglinide,
  + chlorpropamide,
  + glimepiride,
  + acetohexamide,
  + glipizide,
  + glyburide,
  + tolbutamide,
  + pioglitazone,
  + rosiglitazone,
  + acarbose,
  + miglitol,
  + troglitazone,
  + tolazamide,
  + examide,
  + sitagliptin,
  + insulin,
  + glyburide-metformin,
  + glipizide-metformin,
  + glimepiride-pioglitazone,
  + metformin-rosiglitazone,
  + metformin-pioglitazone,

The feature indicates whether the drug was prescribed or there was a change in the dosage. Values:“up” if the dosage was increased during the encounter, “down” if the dosage was decreased, “steady” if the dosage did not change, and “no” if the drug was not prescribed

* Readmitted: Days to inpatient readmission. Values: “<30” if the patient was readmitted in less than 30 days, “>30” if the patient was readmitted in more than 30 days, and “No” for no record of readmission.

Numeric Data

* Weight: Weight in pounds
* Time in hospital: Integer number of days between admission and discharge
* Number of lab procedures: Number of lab tests performed during the encounter
* Number of procedures: Number of procedures (other than lab tests) performed during the encounter
* Number of medications: Number of distinct generic names administered during the encounter
* Number of outpatient visits: Number of outpatient visits of the patient in the year preceding the encounter
* Number of emergency visits: Number of emergency visits of the patient in the year preceding the encounter
* Number of inpatient visits: Number of inpatient visits of the patient in the year preceding the encounter

A function "mean", is used to find the mean of all the numeric data. On doing this we find that columns "weight","num\_outpatient"," num\_inpatient" and "num\_emergency" are not numeric and in fact are nominal. Ignoring these for the mean function we find the following results.

|  |  |
| --- | --- |
| Variable | Mean |
| num\_lab\_procedures | 43.0786 |
| num\_procedures | 1.3992 |
| num\_medications | 15.5638 |
| number\_diagnoses | 7.0253 |
| time\_in\_hospital | 4.4347 |

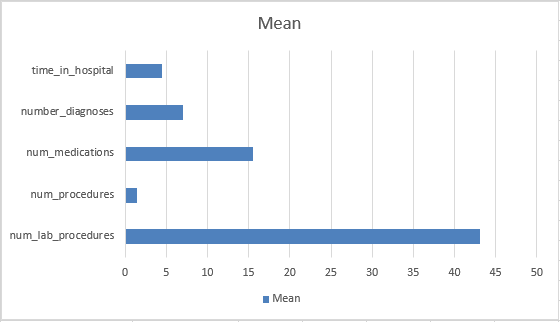
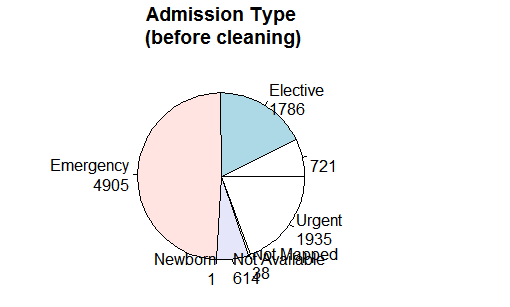


Figure 1:Mean of Numerical values

**DATA CLEANING**

Dirty data is data that is filled with inconsistencies and hence produces unreliable output. The Data cleaning routines work to “clean” the data by ﬁlling in missing values, smoothing noisy data, identifying or removing outliers, and resolving inconsistencies [2].



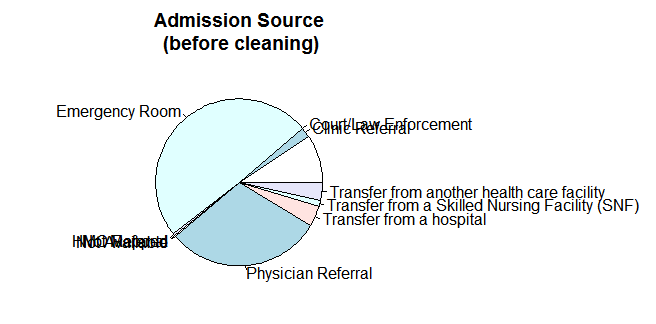


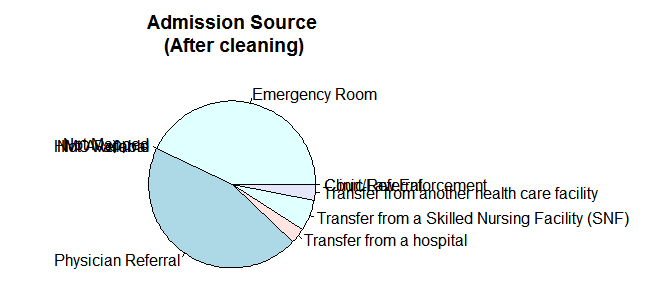
Figure 2:Examples of data before cleaning

As seen, the data columns consists of blanks, "?" And "NA"'s . To go further in our analysis, we have to remove all of this. The process is as follows

* Removal of blanks from Admission Source, Discharge disposition and Admission type variables.
* Replacing "?" Data values with NA in race, weight, payer\_code, diag\_1, diag\_2, diag\_3, and medical\_speciality columns
* Removing NA values

The result of the process is seen below for each of these columns:

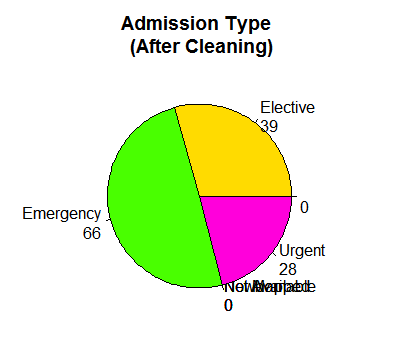
Admission Source



|  |  |
| --- | --- |
| Level | Frequency |
| Clinical Referral | 0 |
| Court/Law Enforcement | 0 |
| Emergency Room | 57 |
| HMO Referral | 0 |
| Not Available | 0 |
| Not Mapped | 0 |
| Physician Referral | 60 |
| Transfer from a hospital | 4 |
| Transfer from a Skilled Nursing Facility (SNF) | 8 |
| Transfer from another health care facility | 4 |

Figure 3: Admission Source column data after cleaning

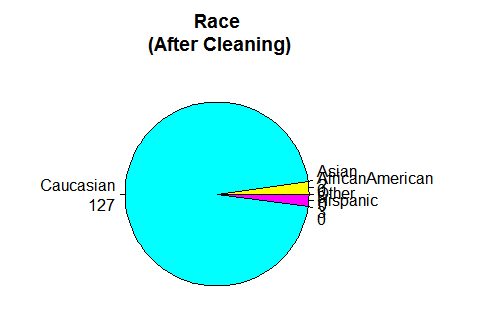
Admission Type



|  |  |
| --- | --- |
| Level | Frequency |
| Emergency | 66 |
| Elective | 39 |
| Urgent | 28 |
| Not Mapped | 0 |
| Not Available | 0 |
| New Born | 0 |

Figure 4: Admission Type column data after cleaning

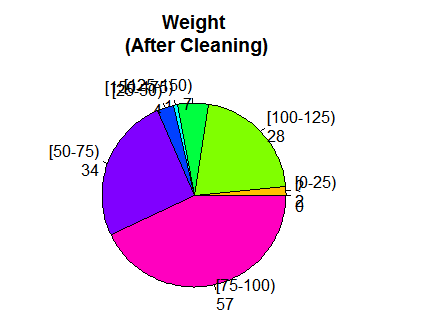
Race



|  |  |
| --- | --- |
| Level | Frequency |
| Caucasian | 127 |
| African American | 3 |
| Hispanic | 0 |
| Asian | 0 |
| Other | 3 |

Figure 5: Race column data after cleaning

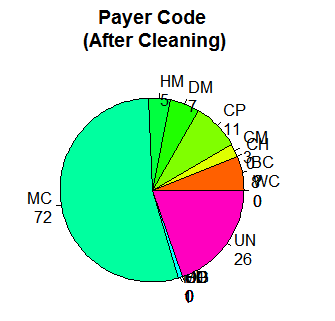
Weight



|  |  |
| --- | --- |
| Level | Frequency |
| [0-25) | 2 |
| [100-125) | 28 |
| [125-150) | 7 |
| [150-175) | 1 |
| [25-50) | 4 |
| [50-75) | 34 |
| [75-100) | 57 |

Figure 6: Weight column data after cleaning

Payer code



|  |  |
| --- | --- |
| Level | Frequency |
| BC | 8 |
| CH | 0 |
| CM | 3 |
| CP | 11 |
| DM | 7 |
| HM | 5 |
| MC | 72 |
| MD | 1 |
| OG | 0 |
| OT | 0 |
| PO | 0 |
| SI | 0 |
| SP | 0 |
| UN | 26 |
| WC | 0 |

Table 1: Payer code column data after cleaning

**CORRELATION**

All tasks are related to readmission. Keeping this in mind, we find out the relation between all other variables and readmission. For the numerical values we perform linear regression against readmission and for all the nominal/categorical data, we perform chi square tests.

**Chi-square**

Two random variables *x*and *y*are called independentif the probability distribution of one variable is not affected by the presence of another. Assume *fij* is the observed frequency count of events belonging to both *i*-th category of *x*and *j*-th category of *y*. Also assume *eij* to be the corresponding expected count if *x*and *y*are independent. The χ2 statistic tests the null hypothesis that x and y are independent, that is, there is no correlation between them. The null hypothesis of the independence assumption is to be rejected if the p-value of the following [Chi-squared](http://www.r-tutor.com/node/61) test statistics is less than a given significance level *α*. [3]

                2
χ2 = ∑  (fij --eij)-
     i, j   eij


The function chisq.test performs chi-squared contingency table tests and goodness-of-fit tests. However the data involved should be binary, i.e. no more than two levels. To overcome this issue, we used the function xtab. Xtab creates a contingency table (optionally a sparse matrix) from cross-classifying factors, usually contained in a data frame, using a formula interface. [4]

Eg:

https://lh6.googleusercontent.com/pw5RSC41zslQtJ85N3yYMguUGuszR-519oSqO0q1mrFgMAt4uFB3hUVmajlGwhtTAAQCwNvUTO7AJN-8F2jkVSEW8KEnb1OnGxWL_eXAeh9QSuE3Ogp9wCkCJrDUSEhqV0zhd_El

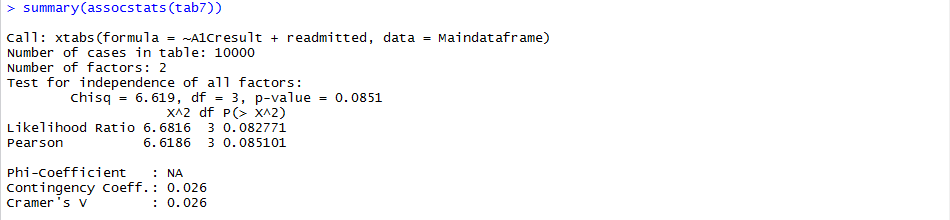
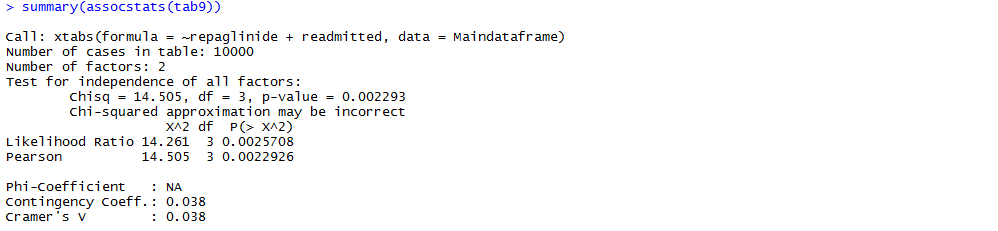


Figure 3: Example of xtab() results

To summarize the data found, we need a function called assocstats() that belongs to the package “vcd”.

Using the xtab(), we found the chi-square value or rather the dependence value between all the nominal variables and readmission. To have a relation of significance, the chi-square value has to be high and the p-value less.

Some of the results of the chi-square tests are shown below



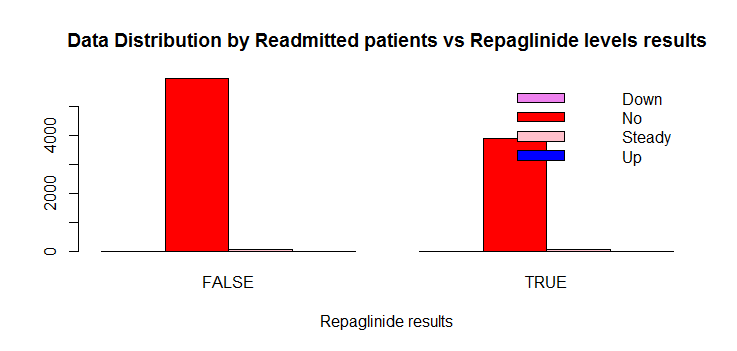
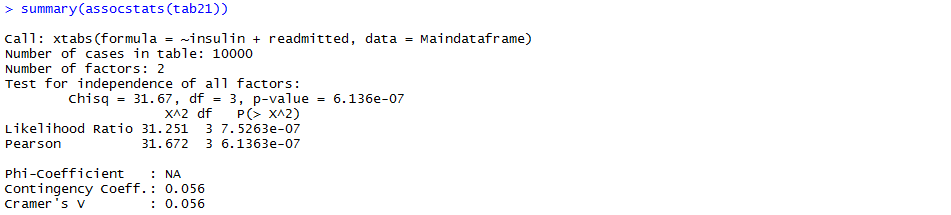


Figure 4:Chi-Square result of Repaglinide against Readmission



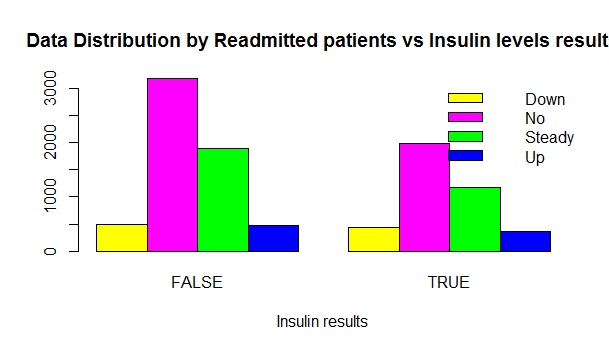
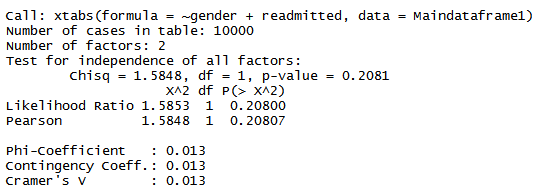


Figure 5: Chi-Square result of Insulin against Readmission



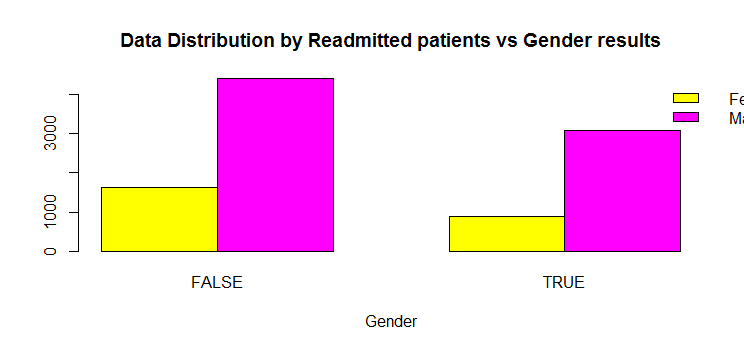
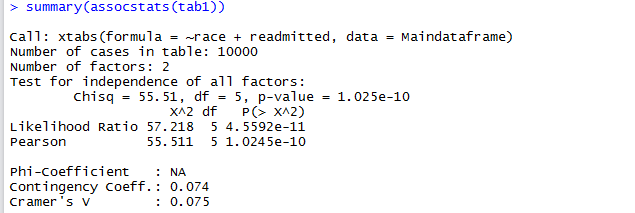


Figure 6: Chi-Square result of Gender against Readmission



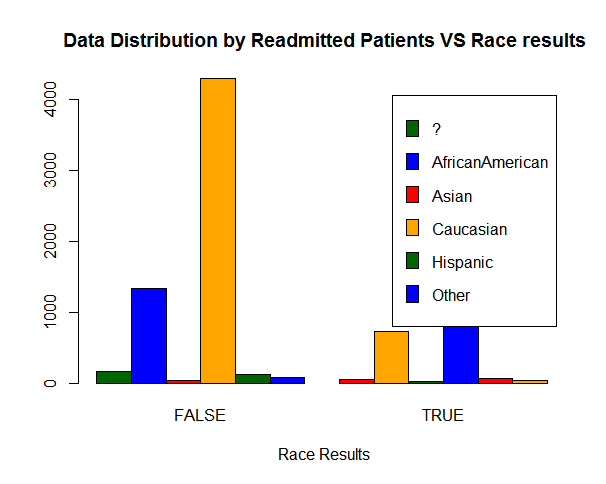


Figure 7: Chi-Square result of Race against Readmission

Based on these results, we considered the following variables (from different types of data such as generic, medicines etc.)

|  |  |
| --- | --- |
| Variable | Chi-Square Value |
| Race | 55.51 |
| Admission\_type\_id | 61.39 |
| Medical\_speciality | 14.6 |
| DiabetesMed | 28.281 |
| Insulin | 31.67 |
| Pioglitazone | 30.29 |
| Repaglinide | 14.50 |
| Change | 26.07 |

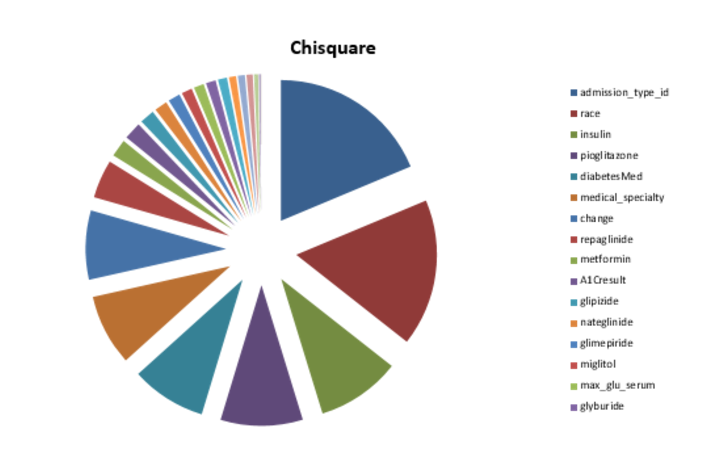


Figure 8: Chi-Square result comparison

**Linear Regression**

One of the most frequent used techniques in statistics is linear regression where we investigate the potential relationship between a variable of interest (often called the response variable but there are many other names in use) and a set of one of more variables (known as the independent variables or some other term). Unsurprisingly there are flexible facilities in **R** for fitting a range of linear models from the simple case of a single variable to more complex relationships.[5]

We used the lm() function. lm is used to fit linear models. It can be used to carry out regression, single stratum analysis of variance and analysis of covariance.

The result of linear regression is as follows

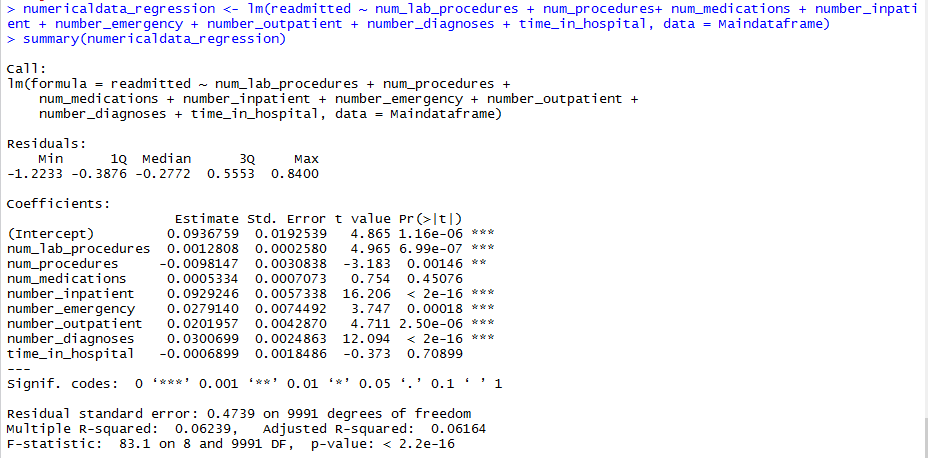


Figure 9: Linear Regression Result

Based on the above data we can say that number\_inpatient, number\_diagnoses and num\_lab\_procedures have the highest relation with readmission.

**PRINCIPAL COMPONENT ANALYSIS**

PCA is a dimension reduction method that searches for best vectors that can be used to represent the data. After running PCA on all numerical attributes, we found that four dimensions were the best for representing the data, namely, (1) Number of Lab Procedures (2) Number of Medications , (3) Number of Diagnoses, and (4) Number of procedures. Then, we run it again to acquire the analysis results for these four attributes, as we show next.

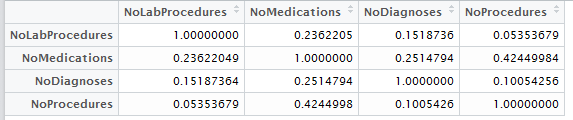


Figure 10: R excerpt of PCA numerical attributes

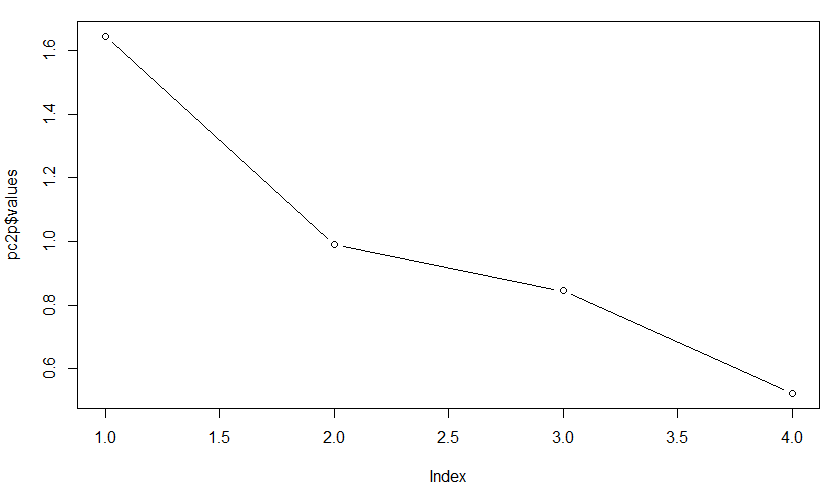


Figure 11: R experpt of Index-values PCA analysis

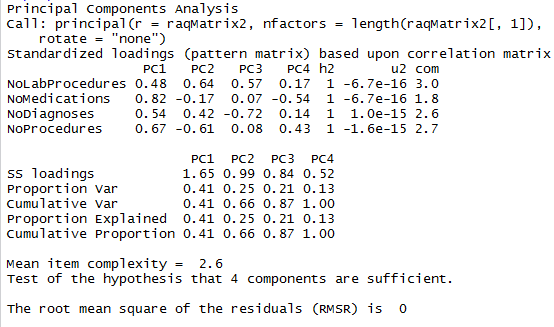


Figure 12: PCA analysis MIC and RMSR

**Additional Prediction TASKS**

**Task 1: Bayesian Network**

A Bayesian network is a probabilistic [graphical model](https://en.wikipedia.org/wiki/Graphical_model) (a type of [statistical model](https://en.wikipedia.org/wiki/Statistical_model)) that represents a set of [random variables](https://en.wikipedia.org/wiki/Random_variables) and their [conditional dependencies](https://en.wikipedia.org/wiki/Conditional_independence) via a [directed acyclic graph](https://en.wikipedia.org/wiki/Directed_acyclic_graph) (DAG). [6].

The incentive behind using Bayesian network is the assumption that attributes are not independent. This type of network will show the dependency between variables in the form of probability shown by an edge that is only shown based on having probability above a certain threshold (th = 0.2).

Using the chi-square tests we selected the variables that have the highest correlation to readmission processes (See: Figure 13) and created a Bayesian network out of it.

**bnlearn** is an R package for learning the graphical structure of Bayesian networks, estimate their parameters and perform some useful inference [7]. There are two types of Bayesian networks: Constraint Based and Score based. We chose to use a Score based network and for this we used the algorithm of Hill Climbing.

bn.hc <- hc ( <name of dataframe>, score = "")

The data frame should consist of only those columns whose network we want to create. For categorical variables there are different kinds of scores such as the Akaike Information Criterion score (aic), the Bayesian Information Criterion score (bic), the logarithm of the Bayesian Dirichlet equivalent score (bde). [8]

We chose the Bayesian Information Criterion (bic) score which is the default score. Each score provides a different threshold of likeliness.

We created a data frame of the variables and the network created is as follows

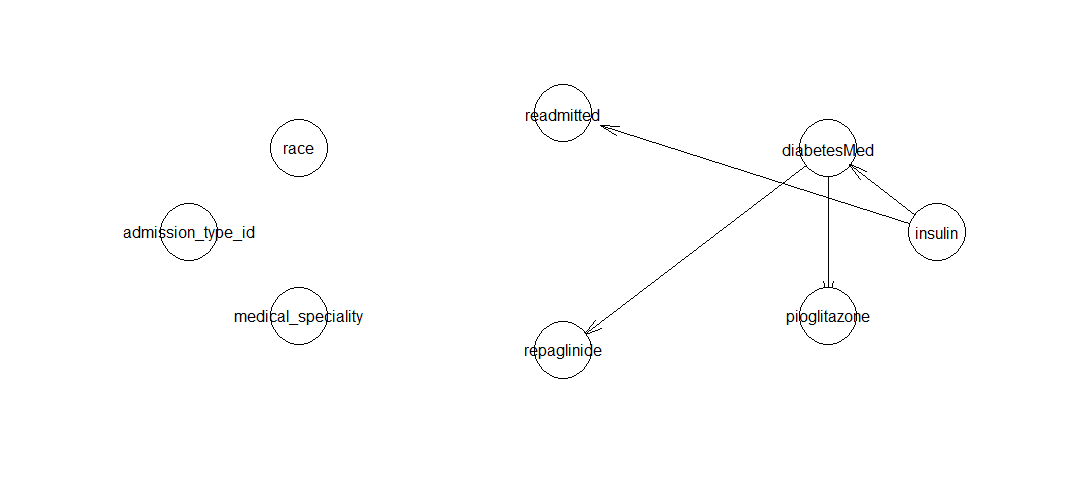


Figure 13:Bayesian Network of highly correlated variables

This image shows to us that while individually each of the variables may seem to have a high correlation with readmission, when combined together only insulin has an effect on readmission. This can be better explained with the following example.

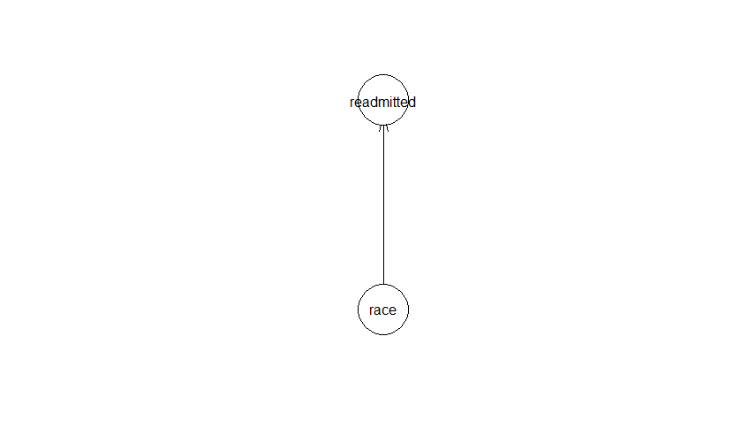
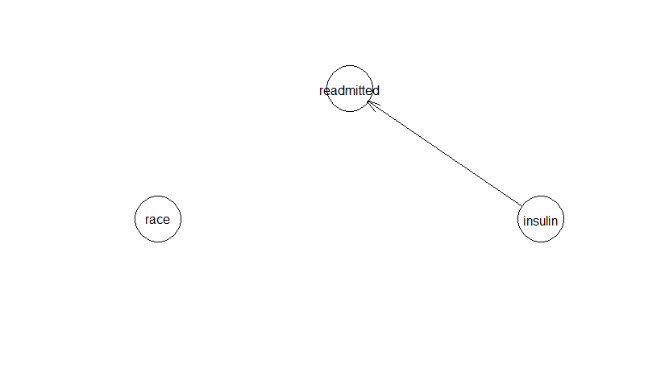
 

Figure: Example of Bayesian Network working

On its own, race has an influence on readmission. However when we consider insulin, we find that insulin has more influence on readmission, while race has none on readmission as well as insulin.

Thus we tried to find bayesian network that can be created of the entire data.

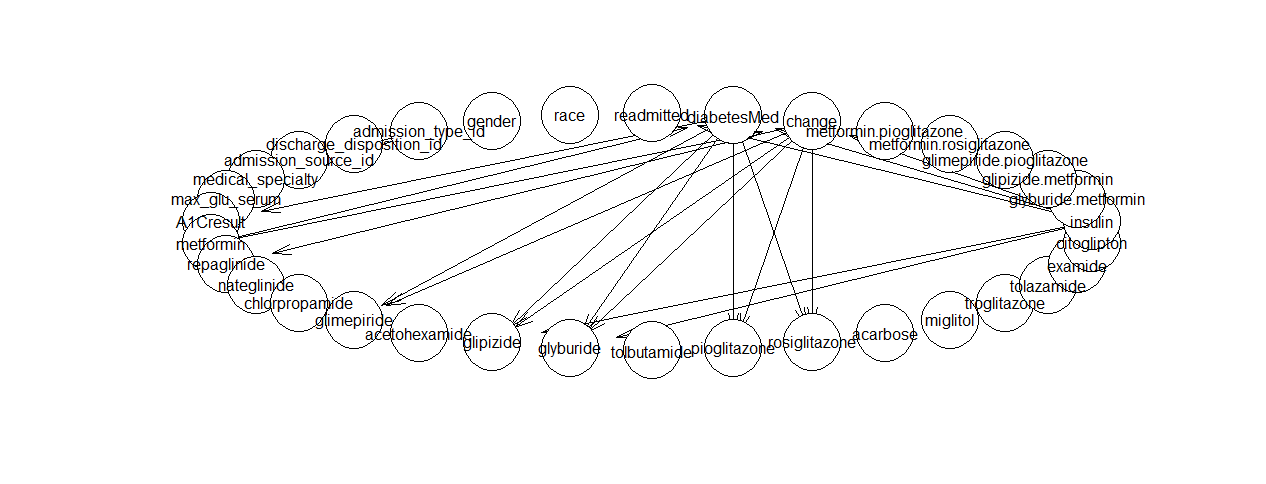


Figure 14:Bayesian Network of all categorical variables

As it is seen, there are a lot of variables that have no relevance to readmission. Most of them are medicines. So we created a network with just the medicines that may have an effect, change in medication, diabetesMed and readmission.

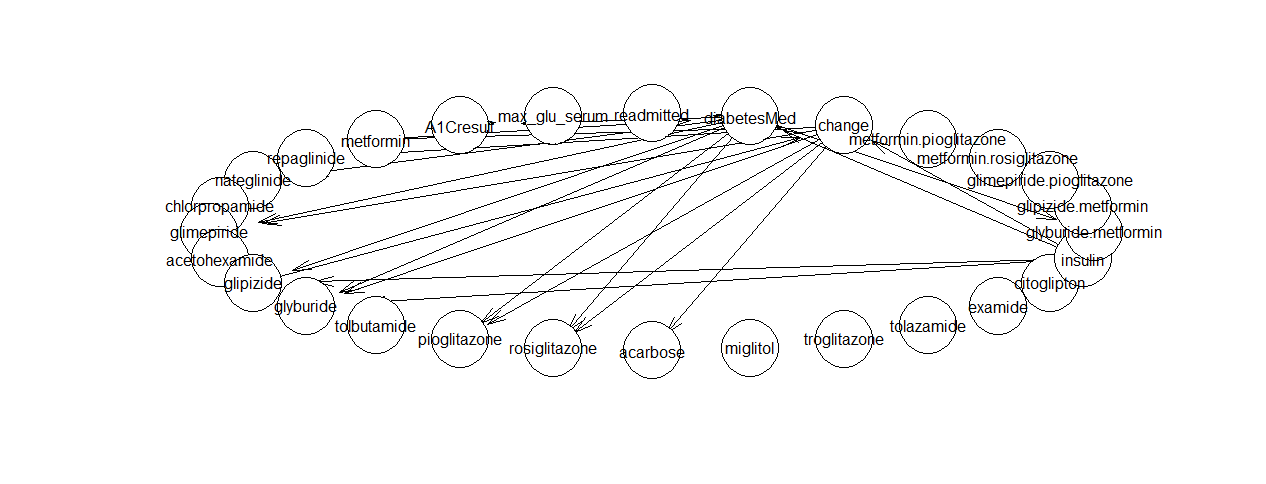


Figure 15: Bayesian Network of all medicine related variables

Again we remove the unnecessary variables, and get the following Bayesian Network.

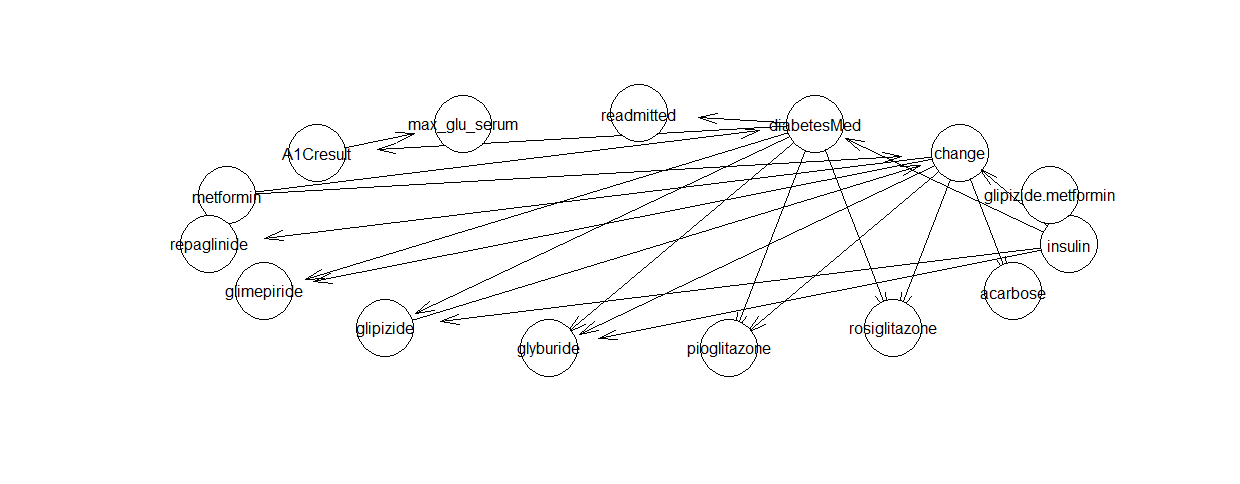


Figure 16: Bayesian Network

As seen in the above Bayesian Networks, readmission is influenced by diabetesMed which in turn is affected by insulin.

To find the individual probabilities of various joint variables, we use the function bn.fit() wherein we enter the Bayesian network and the data frame from where it was created

Eg: fittedbn <- bn.fit(hc(trial10), trial10)

Note: the readmission column was changed to 0 if FALSE and 1 if TRUE

**Task 2 – Hazard Function**

Hazard function is a well known function that is used to estimate the probability of expire for patients based on several attributes such as time spent in ICU (we don’t have such attribute in this dataset) . Since our data does not contain such attribute, we used the following three numeric attribute as a duration attribute. This assumption states that following attributes, indicate the time spent in hospital as more procedures and diagnoses mean more time spent at the hospital:

* + Number of lab procedures
  + Number of procedures
  + Number of diagnoses
* For this purpose, we utilize a variation of Hazard function that is **Cox regression** (or **proportional hazards regression**).

The following snippet of code shows our usage Cox regression function:

res.cox1 <- coxph(trainData$SurvObj ~ trainData$num\_lab\_procedures + trainData$num\_procedures + trainData$number\_diagnoses, data = trainData)

We split the data into training and testing set, 80% and 20% respectively and the ROC function to estimate the accuracy if our work was estimated to be %58 ROC. This estimation is shown by the following snippet of code:

roc(testData$readmitted, pred1, data= "testData",smooth=TRUE)

Note: Hazard function is usually used to predict mortality (expiry) for patients. We treat paitent readmission flag (readmitted or not) as a true or false value in order to use this function.

**Task 3: Text Analysis**

Each patient has a diagnosis description attribute (3 attributes in fact). We combine these three attributes together as one diagnosis and we do the following text processing based on this aggregation.

**Preprocessing:** we remove all stop words and dynamic stop words. Stop words are the ones that do not contribute to the meaning of the diagnosis such as the, a, an, etc. Dynamic stop words are the ones that are mentioned by many documents so they cannot be used to distinguish between documents. For example, using the feature “Grad student” to distinguish between people in the graduate school would not be a good feature. The threshold we used for this experiment is (th = 0.4), intuitively, this means if any word is mentioned by more than 40% of documents (patient records) then it will be filtered out.

**Knowledge graph:** Some words in the diagnosis are related or the same. Using them as different words might be the best approach to measure the similarity between patients based on their diagnosis description. For instance, **tachycardia** is a form (or a subset) of **heart disease.** Using classical text similarity measures such as (tf-idf) is not the best approach due to the mention similarity. For this purpose, we use a medical ontology DIODE [9] to extract related terms if they are either medical variations of the same term.

**WordNet:** WordNet provides ontology for English language that provides relations between terms such as (speak and whisper). These relationships are of a hieratical fashion that go from general to specific. For instance, communicate is term that is a node in the ontology, this node a child (speak), speak in its turn has a child that is (whisper). By traversing this tree, synonyms, subsets, and different forms of the same word could be collected. However, the more distance (number of edges or nodes) between any two terms in this tree the less similar they are.

In this work, we calculate the distance between any two terms by counting the number of nodes. If this number is less than a predefined threshold, then we consider the two words to be the same word. In our experiments, this threshold was set to be 4 (i.e., if two terms have only four nodes between them, then they are considered to be one term)

We tried different prediction algorithms and the best one that yielded the highest accuracy was RandomForest as it yielded 63% accuracy to predict patient readmission.

**CONCLUSION**

In this work, we have experimented with the provided dataset and explored various prediction algorithms after performing cleaning, preprocessing, and dimension reduction. Based on our findings using an enhanced text analysis approach utilizing knowledge graphs and WordNet English ontology, in addition to four numerical attributes acquired by performing PCA dimension reduction, is the best approach to predict patient readmission. Patient mortality can also be predicted based on the same synthesized feature space and we keep this part for future use.

**References:**

[1] Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records – Beata Strack, Jonathan P.DeShazo,Chris Gennings, Juan L. Olmo, Sebastian Ventura, Krzysztof J. Cios, and John N. Clore

[2] Data Mining Concepts and Techniques, Third Edition, Authors - Jiawei Han, Micheline Kamber, Jian Pei

[3] <http://www.r-tutor.com/elementary-statistics/goodness-fit/chi-squared-test-independence>

[4] <https://stat.ethz.ch/R-manual/R-devel/library/stats/html/xtabs.html>

[5] <https://www.r-bloggers.com/simple-linear-regression-2/>

[6] <https://en.wikipedia.org/wiki/Bayesian_network>

[7] <http://www.bnlearn.com/>

[8] <https://cran.r-project.org/web/packages/bnlearn/bnlearn.pdf>

[9] Lee, Yugyung, and James Geller. "Semantic enrichment for medical ontologies." *Journal of biomedical informatics* 39.2 (2006): 209-226.

R Code - A snippet

> setwd("C:/Users/Tanya Peddi/Documents/UNCC Academic/6162 Knowledge Discovery in Databases/Datasets/Main project")

> dataframe <- read.csv("10kDiabetes.csv", header = TRUE)

> Maindataframe1 <- data.frame(dataframe)

> View(Maindataframe1)

> mean(Maindataframe1[["num\_lab\_procedures"]])

[1] 43.0786

> mean(Maindataframe1[["num\_procedures"]])

[1] 1.3992

> mean(Maindataframe1[["num\_medications"]])

[1] 15.5638

> mean(Maindataframe1[["num\_outpatient"]])

[1] NA

Warning message:

In mean.default(Maindataframe1[["num\_outpatient"]]) :

argument is not numeric or logical: returning NA

> mean(Maindataframe1[["num\_inpatient"]])

[1] NA

Warning message:

In mean.default(Maindataframe1[["num\_inpatient"]]) :

argument is not numeric or logical: returning NA

> mean(Maindataframe1[["num\_emergency"]])

[1] NA

Warning message:

In mean.default(Maindataframe1[["num\_emergency"]]) :

argument is not numeric or logical: returning NA

> mean(Maindataframe1[["number\_diagnoses"]])

[1] 7.0253

> mean(Maindataframe1[["weight"]])

[1] NA

Warning message:

In mean.default(Maindataframe1[["weight"]]) :

argument is not numeric or logical: returning NA

>

>

> Maindataframe1$admission\_type\_id <- replace(Maindataframe1$admission\_type\_id,Maindataframe1$admission\_type\_id =='',NA)

>

> Maindataframe1$discharge\_disposition\_id <- replace(Maindataframe1$discharge\_disposition\_id,Maindataframe1$discharge\_disposition\_id=='',NA)

>

> mytable <- table(Maindataframe$admission\_type\_id)

> lbls <- paste(names(mytable), "\n", mytable, sep="")

> pie(mytable, labels = lbls,

+ main="Admission Type\n (before cleaning)")

> mytable <- table(Maindataframe$discharge\_disposition\_id)

> lbls <- paste(names(mytable), "\n", mytable, sep="")

> pie(mytable, labels = lbls,

+ main="Discharge Disposition\n (before cleaning)")

> mytable <- table(Maindataframe$discharge\_disposition\_id)

> lbls <- paste(names(mytable), "\n", mytable, sep="")

> pie(mytable,

+ main="Discharge Disposition\n (before cleaning)")

> mytable <- table(Maindataframe$admission\_source\_id)

> lbls <- paste(names(mytable), "\n", mytable, sep="")

> pie(mytable,

+ main="Admission Source\n (before cleaning)")

>

>

>

> Maindataframe1$race<-replace(Maindataframe1$race,Maindataframe1$race=='?',NA)

> Maindataframe1$weight<-replace(Maindataframe1$weight,Maindataframe1$weight=='?',NA)

> Maindataframe1$payer\_code<-replace(Maindataframe1$payer\_code,Maindataframe1$payer\_code=='?',NA)

> Maindataframe1$diag\_1<-replace(Maindataframe1$diag\_1,Maindataframe1$diag\_1=='?',NA)

> Maindataframe1$diag\_2<-replace(Maindataframe1$diag\_2,Maindataframe1$diag\_2=='?',NA)

> Maindataframe1$diag\_3<-replace(Maindataframe1$diag\_3,Maindataframe1$diag\_3=='?',NA)

> Maindataframe1$medical\_specialty<-replace(Maindataframe1$medical\_specialty,Maindataframe1$medical\_specialty=='?',NA)

> Maindataframe1$medical\_specialty<-replace(Maindataframe1$medical\_specialty,Maindataframe1$medical\_specialty=='?',NA)

> summary(Maindataframe1)

rowID race gender age

Min. : 1 ? : 0 Female:5398 [70-80):2595

1st Qu.: 2501 AfricanAmerican:2063 Male :4602 [60-70):2187

Median : 5000 Asian : 55 [50-60):1722

Mean : 5000 Caucasian :7359 [80-90):1577

3rd Qu.: 7500 Hispanic : 181 [40-50): 978

Max. :10000 Other : 121 [30-40): 417

NA's : 221 (Other): 524

weight admission\_type\_id

[75-100) : 176 Emergency :4905

[50-75) : 116 Urgent :1935

[100-125): 76 Elective :1786

[125-150): 18 Not Available: 614

[25-50) : 12 Not Mapped : 38

(Other) : 10 (Other) : 1

NA's :9592 NA's : 721

discharge\_disposition\_id

Discharged to home :6056

Discharged/transferred to SNF :1190

Discharged/transferred to home with home health service:1160

Expired : 195

Discharged/transferred to another short term hospital : 185

(Other) : 745

NA's : 469

admission\_source\_id time\_in\_hospital payer\_code

Emergency Room :4940 Min. : 1.000 MC :2519

Physician Referral :3010 1st Qu.: 2.000 HM : 459

: 936 Median : 4.000 BC : 383

Transfer from a hospital : 396 Mean : 4.435 SP : 303

Transfer from another health care facility: 349 3rd Qu.: 6.000 UN : 284

Clinic Referral : 180 Max. :14.000 (Other): 711

(Other) : 189 NA's :5341

medical\_specialty num\_lab\_procedures num\_procedures num\_medications

InternalMedicine :1921 Min. : 1.00 Min. :0.000 Min. : 1.00

Family/GeneralPractice: 862 1st Qu.: 32.00 1st Qu.:0.000 1st Qu.:10.00

Cardiology : 698 Median : 44.00 Median :1.000 Median :14.00

Emergency/Trauma : 497 Mean : 43.08 Mean :1.399 Mean :15.56

Surgery-General : 333 3rd Qu.: 57.00 3rd Qu.:2.000 3rd Qu.:19.00

(Other) :1589 Max. :120.00 Max. :6.000 Max. :81.00

NA's :4100

number\_outpatient number\_emergency number\_inpatient diag\_1 diag\_2

Min. : 0.0000 Min. : 0.000 Min. : 0.0000 414 : 735 250 : 684

1st Qu.: 0.0000 1st Qu.: 0.000 1st Qu.: 0.0000 428 : 636 276 : 658

Median : 0.0000 Median : 0.000 Median : 0.0000 786 : 454 428 : 637

Mean : 0.2817 Mean : 0.115 Mean : 0.3873 410 : 403 427 : 535

3rd Qu.: 0.0000 3rd Qu.: 0.000 3rd Qu.: 0.0000 486 : 312 401 : 402

Max. :36.0000 Max. :42.000 Max. :10.0000 (Other):7458 (Other):7025

NA's : 2 NA's : 59

diag\_3 number\_diagnoses max\_glu\_serum A1Cresult metformin

250 :1276 Min. :1.000 >200: 197 >7 : 353 Down : 51

401 : 962 1st Qu.:5.000 >300: 126 >8 : 820 No :8011

276 : 448 Median :7.000 None:9336 None:8379 Steady:1816

428 : 415 Mean :7.025 Norm: 341 Norm: 448 Up : 122

427 : 392 3rd Qu.:9.000

(Other):6299 Max. :9.000

NA's : 208

repaglinide nateglinide chlorpropamide glimepiride acetohexamide

Down : 5 Down : 1 No :9987 Down : 9 No:10000

No :9870 No :9949 Steady: 12 No :9509

Steady: 112 Steady: 49 Up : 1 Steady: 450

Up : 13 Up : 1 Up : 32

glipizide glyburide tolbutamide pioglitazone rosiglitazone acarbose

Down : 52 Down : 61 No :9997 Down : 13 Down : 9 No :9968

No :8696 No :8814 Steady: 3 No :9294 No :9239 Steady: 31

Steady:1160 Steady:1033 Steady: 664 Steady: 730 Up : 1

Up : 92 Up : 92 Up : 29 Up : 22

miglitol troglitazone tolazamide examide citoglipton insulin

Down : 1 No:10000 No :9996 No:10000 No:10000 Down : 942

No :9995 Steady: 4 No :5159

Steady: 3 Steady:3057

Up : 1 Up : 842

glyburide.metformin glipizide.metformin glimepiride.pioglitazone

Down : 1 No :9998 No:10000

No :9944 Steady: 2

Steady: 53

Up : 2

metformin.rosiglitazone metformin.pioglitazone change diabetesMed

No:10000 No:10000 Ch:4276 No :2522

No:5724 Yes:7478

readmitted

Mode :logical

FALSE:6035

TRUE :3965

NA's :0

diag\_1\_desc

Coronary atherosclerosis of unspecified type of vessel, native or graft : 735

Congestive heart failure, unspecified : 636

Respiratory abnormality, unspecified : 454

Acute myocardial infarction of anterolateral wall, episode of care unspecified: 403

Pneumonia, organism unspecified : 312

Paroxysmal supraventricular tachycardia : 279

(Other) :7181

diag\_2\_desc

Diabetes mellitus without mention of complication, type II or unspecified type, not stated as uncontrolled: 684

Hyperosmolality and/or hypernatremia : 658

Congestive heart failure, unspecified : 637

Paroxysmal supraventricular tachycardia : 535

Malignant essential hypertension : 402

Chronic airway obstruction, not elsewhere classified : 357

(Other) :6727

diag\_3\_desc

Diabetes mellitus without mention of complication, type II or unspecified type, not stated as uncontrolled:1276

Malignant essential hypertension : 962

Hyperosmolality and/or hypernatremia : 448

Congestive heart failure, unspecified : 415

Paroxysmal supraventricular tachycardia : 392

Coronary atherosclerosis of unspecified type of vessel, native or graft : 361

(Other) :6146

>

> library(VIM)

Loading required package: colorspace

Loading required package: grid

Loading required package: data.table

data.table 1.10.4

The fastest way to learn (by data.table authors): https://www.datacamp.com/courses/data-analysis-the-data-table-way

Documentation: ?data.table, example(data.table) and browseVignettes("data.table")

Release notes, videos and slides: http://r-datatable.com

VIM is ready to use.

Since version 4.0.0 the GUI is in its own package VIMGUI.

Please use the package to use the new (and old) GUI.

Suggestions and bug-reports can be submitted at: https://github.com/alexkowa/VIM/issues

Attaching package: ‘VIM’

The following object is masked from ‘package:datasets’:

sleep

Warning messages:

1: In doTryCatch(return(expr), name, parentenv, handler) :

display list redraw incomplete

2: In doTryCatch(return(expr), name, parentenv, handler) :

invalid graphics state

3: In doTryCatch(return(expr), name, parentenv, handler) :

invalid graphics state

4: In doTryCatch(return(expr), name, parentenv, handler) :

display list redraw incomplete

5: In doTryCatch(return(expr), name, parentenv, handler) :

invalid graphics state

6: In doTryCatch(return(expr), name, parentenv, handler) :

invalid graphics state

> DiabeticData1<-kNN(DiabeticData,variable=c("race","admission\_type\_id","medical\_specialty","payer\_code","weight","diag\_1","diag\_2","diag\_3"),k=10)

Error in kNN(DiabeticData, variable = c("race", "admission\_type\_id", "medical\_specialty", :

object 'DiabeticData' not found

>

>

> Maindataframe2<- na.omit(Maindataframe1)

>

> summary(Maindataframe2)

rowID race gender age weight admission\_type\_id

Min. : 120 ? : 0 Female:67 [70-80):34 [75-100) :57 : 0

1st Qu.:2174 AfricanAmerican: 3 Male :66 [80-90):31 [50-75) :34 Elective :39

Median :4753 Asian : 0 [60-70):29 [100-125):28 Emergency :66

Mean :4756 Caucasian :127 [50-60):22 [125-150): 7 Newborn : 0

3rd Qu.:7220 Hispanic : 0 [40-50):10 [25-50) : 4 Not Available: 0

Max. :9972 Other : 3 [20-30): 3 [0-25) : 2 Not Mapped : 0

(Other): 4 (Other) : 1 Urgent :28

discharge\_disposition\_id admission\_source\_id

Discharged to home :93 Physician Referral :60

Discharged/transferred to home with home health service :23 Emergency Room :57

Discharged/transferred to SNF :11 Transfer from a Skilled Nursing Facility (SNF): 8

Discharged/transferred to another type of inpatient care institution: 2 Transfer from a hospital : 4

Hospice / home : 2 Transfer from another health care facility : 4

Discharged/transferred to another short term hospital : 1 : 0

(Other) : 1 (Other) : 0

time\_in\_hospital payer\_code medical\_specialty num\_lab\_procedures num\_procedures num\_medications number\_outpatient number\_emergency

Min. : 1.000 MC :72 Cardiology :45 Min. : 1.00 Min. :0.000 Min. : 1.00 Min. : 0.000 Min. :0.0000

1st Qu.: 2.000 UN :26 InternalMedicine :42 1st Qu.:41.00 1st Qu.:0.000 1st Qu.:10.00 1st Qu.: 0.000 1st Qu.:0.0000

Median : 4.000 CP :11 Surgery-General :32 Median :56.00 Median :1.000 Median :15.00 Median : 1.000 Median :0.0000

Mean : 4.271 BC : 8 Family/GeneralPractice : 9 Mean :52.34 Mean :1.541 Mean :15.19 Mean : 1.617 Mean :0.1654

3rd Qu.: 6.000 DM : 7 Psychiatry : 4 3rd Qu.:68.00 3rd Qu.:2.000 3rd Qu.:19.00 3rd Qu.: 2.000 3rd Qu.:0.0000

Max. :13.000 HM : 5 ObstetricsandGynecology: 1 Max. :86.00 Max. :6.000 Max. :41.00 Max. :14.000 Max. :2.0000

(Other): 4 (Other) : 0

number\_inpatient diag\_1 diag\_2 diag\_3 number\_diagnoses max\_glu\_serum A1Cresult metformin repaglinide nateglinide chlorpropamide

Min. :0.0000 414 :12 428 :14 250 :15 Min. :3.000 >200: 0 >7 : 1 Down : 1 Down : 0 Down : 0 No :133

1st Qu.:0.0000 427 :10 424 :10 401 :10 1st Qu.:7.000 >300: 0 >8 : 1 No :103 No :131 No :133 Steady: 0

Median :0.0000 996 : 9 427 : 8 424 :10 Median :9.000 None:133 None:131 Steady: 26 Steady: 2 Steady: 0 Up : 0

Mean :0.4286 715 : 8 401 : 7 427 :10 Mean :8.158 Norm: 0 Norm: 0 Up : 3 Up : 0 Up : 0

3rd Qu.:1.0000 786 : 6 276 : 5 396 : 8 3rd Qu.:9.000

Max. :4.0000 428 : 4 413 : 5 414 : 5 Max. :9.000

(Other):84 (Other):84 (Other):75

glimepiride acetohexamide glipizide glyburide tolbutamide pioglitazone rosiglitazone acarbose miglitol troglitazone tolazamide examide

Down : 0 No:133 Down : 0 Down : 0 No :133 Down : 0 Down : 0 No :132 Down : 0 No:133 No :133 No:133

No :123 No :107 No :114 Steady: 0 No :119 No :119 Steady: 1 No :133 Steady: 0

Steady: 8 Steady: 19 Steady: 17 Steady: 14 Steady: 14 Up : 0 Steady: 0

Up : 2 Up : 7 Up : 2 Up : 0 Up : 0 Up : 0

citoglipton insulin glyburide.metformin glipizide.metformin glimepiride.pioglitazone metformin.rosiglitazone metformin.pioglitazone change

No:133 Down : 2 Down : 0 No :133 No:133 No:133 No:133 Ch:41

No :118 No :133 Steady: 0 No:92

Steady: 12 Steady: 0

Up : 1 Up : 0

diabetesMed readmitted diag\_1\_desc

No :48 Mode :logical Coronary atherosclerosis of unspecified type of vessel, native or graft :12

Yes:85 FALSE:49 Paroxysmal supraventricular tachycardia :10

TRUE :84 Mechanical complication of unspecified cardiac device, implant, and graft: 9

NA's :0 Osteoarthrosis, generalized, site unspecified : 8

Respiratory abnormality, unspecified : 6

Congestive heart failure, unspecified : 4

(Other) :84

diag\_2\_desc

Congestive heart failure, unspecified :14

Mitral valve disorders :10

Paroxysmal supraventricular tachycardia : 8

Malignant essential hypertension : 7

Angina decubitus : 5

Coronary atherosclerosis of unspecified type of vessel, native or graft: 5

(Other) :84

diag\_3\_desc

Diabetes mellitus without mention of complication, type II or unspecified type, not stated as uncontrolled:15

Malignant essential hypertension :10

Mitral valve disorders :10

Paroxysmal supraventricular tachycardia :10

Mitral valve stenosis and aortic valve stenosis : 8

Coronary atherosclerosis of unspecified type of vessel, native or graft : 5

(Other) :75

> mytable <- table(Maindataframe2$admission\_source\_id)

> lbls <- paste(names(mytable), "\n", mytable, sep="")

> pie(mytable,

+ main="Admission Source\n (After cleaning)")

Error in plot.new() : figure margins too large

> mytable <- table(Maindataframe2$admission\_source\_id)

> lbls <- paste(names(mytable), "\n", mytable, sep="")

> pie(mytable,

+ main="Admission Source\n (After cleaning)")

> summary(Maindataframe2$admission\_source\_id)

Clinic Referral

0 0

Court/Law Enforcement Emergency Room

0 57

HMO Referral Not Available

0 0

Not Mapped Physician Referral

0 60

Transfer from a hospital Transfer from a Skilled Nursing Facility (SNF)

4 8

Transfer from another health care facility

4

> mytable <- table(Maindataframe2$admission\_type\_id)

> lbls <- paste(names(mytable), "\n", mytable, sep="")

> pie(mytable,

+ main="Admission Type\n (After cleaning)")

> mytable <- table(Maindataframe2$admission\_type\_id)

> lbls <- paste(names(mytable), "\n", mytable, sep="")

> pie(mytable,labels = lbls, col=rainbow(length(lbls)),

+ main="Admission Type \n (After Cleaning)")

>

>

> library(plotrix)

Error in library(plotrix) : there is no package called ‘plotrix’

> install.packages("plotrix")

Installing package into ‘C:/Users/Tanya Peddi/Documents/R/win-library/3.3’

(as ‘lib’ is unspecified)

There is a binary version available but the source version is later:

binary source needs\_compilation

plotrix 3.6-4 3.6-5 FALSE

installing the source package ‘plotrix’

trying URL 'https://cran.rstudio.com/src/contrib/plotrix\_3.6-5.tar.gz'

Content type 'application/x-gzip' length 237647 bytes (232 KB)

downloaded 232 KB

\* installing \*source\* package 'plotrix' ...

\*\* package 'plotrix' successfully unpacked and MD5 sums checked

\*\* R

\*\* data

\*\* demo

\*\* inst

\*\* preparing package for lazy loading

\*\* help

\*\*\* installing help indices

\*\* building package indices

\*\* testing if installed package can be loaded

\*\*\* arch - i386

\*\*\* arch - x64

\* DONE (plotrix)

The downloaded source packages are in

‘C:\Users\Tanya Peddi\AppData\Local\Temp\RtmpAx2OdD\downloaded\_packages’

> library(plotrix)

> mytable <- table(Maindataframe2$admission\_type\_id)

> lbls <- paste(names(mytable), "\n", mytable, sep="")

> pie(mytable,labels = lbls, col=rainbow(length(lbls)),explode=0.1,

+ main="Admission Type \n (After Cleaning)")

Warning messages:

1: In text.default(1.1 \* P$x, 1.1 \* P$y, labels[i], xpd = TRUE, adj = ifelse(P$x < :

"explode" is not a graphical parameter

2: In text.default(1.1 \* P$x, 1.1 \* P$y, labels[i], xpd = TRUE, adj = ifelse(P$x < :

"explode" is not a graphical parameter

3: In text.default(1.1 \* P$x, 1.1 \* P$y, labels[i], xpd = TRUE, adj = ifelse(P$x < :

"explode" is not a graphical parameter

4: In text.default(1.1 \* P$x, 1.1 \* P$y, labels[i], xpd = TRUE, adj = ifelse(P$x < :

"explode" is not a graphical parameter

5: In text.default(1.1 \* P$x, 1.1 \* P$y, labels[i], xpd = TRUE, adj = ifelse(P$x < :

"explode" is not a graphical parameter

6: In text.default(1.1 \* P$x, 1.1 \* P$y, labels[i], xpd = TRUE, adj = ifelse(P$x < :

"explode" is not a graphical parameter

7: In text.default(1.1 \* P$x, 1.1 \* P$y, labels[i], xpd = TRUE, adj = ifelse(P$x < :

"explode" is not a graphical parameter

8: In title(main = main, ...) : "explode" is not a graphical parameter

> pie3D(mytable,labels=lbls,explode=0.1,

+ main="Admission Type \n (After Cleaning) ")

Error in if (length(by) && by == 0 && length(del) && del == 0) return(from) :

missing value where TRUE/FALSE needed

In addition: There were 16 warnings (use warnings() to see them)

> mytable <- table(Maindataframe2$admission\_type\_id)

> lbls <- paste(names(mytable), "\n", mytable, sep="")

> pie(mytable,labels = lbls, col=rainbow(length(lbls)),

+ main="Admission Type \n (After Cleaning)")

> summary(Maindataframe2$admission\_type\_id)

Elective Emergency Newborn Not Available Not Mapped

0 39 66 0 0 0

Urgent

28

> mytable <- table(Maindataframe2$discharge\_disposition\_id)

> lbls <- paste(names(mytable), "\n", mytable, sep="")

> pie(mytable,labels = lbls, col=rainbow(length(lbls)),

+ main="Discharge Disposition\n (After Cleaning)")

> mytable <- table(Maindataframe2$discharge\_disposition\_id)

> lbls <- paste(names(mytable), "\n", mytable, sep="")

> barplot(mytable, main="Discharge Disposition\n (After Cleaning)",

+ xlab="Discharge Disposition types")

>

> mytable <- table(Maindataframe2$discharge\_disposition\_id)

> lbls <- paste(names(mytable), "\n", mytable, sep="")

> pie(mytable,labels = lbls, col=rainbow(length(lbls)),

+ main="Discharge Disposition\n (After Cleaning)")

> mytable <- table(Maindataframe2$race)

> lbls <- paste(names(mytable), "\n", mytable, sep="")

> pie(mytable,labels = lbls, col=rainbow(length(lbls)),

+ main="Race\n (After Cleaning)")

> mytable <- table(Maindataframe2$race)

> lbls <- paste(names(mytable), "\n", mytable, sep="")

> barplot(mytable, main="Race\n (After Cleaning)",

+ xlab="race types")

>

> mytable <- table(Maindataframe2$race)

> lbls <- paste(names(mytable), "\n", mytable, sep="")

> pie(mytable,labels = lbls, col=rainbow(length(lbls)),

+ main="Race\n (After Cleaning)")

> summary(Maindataframe2$race)

? AfricanAmerican Asian Caucasian Hispanic Other

0 3 0 127 0 3

> mytable <- table(Maindataframe2$diag\_1)

> lbls <- paste(names(mytable), "\n", mytable, sep="")

> pie(mytable,labels = lbls, col=rainbow(length(lbls)),

+ main="Diagnoses 1\n (After Cleaning)")

> summary(Maindataframe2$diag\_1)

414 427 996 715 786 428 491 296 402 410 433 569

12 10 9 8 6 4 4 3 3 3 3 3

578 276 435 560 577 724 780 998 153 198 204 225

3 2 2 2 2 2 2 2 1 1 1 1

230 241 246 250.11 250.22 250.3 250.6 250.8 280 284 285 295

1 1 1 1 1 1 1 1 1 1 1 1

38 426 430 434 438 465 482 486 493 511 53 530

1 1 1 1 1 1 1 1 1 1 1 1

531 532 537 562 574 576 584 590 618 682 722 728

1 1 1 1 1 1 1 1 1 1 1 1

789 8 820 852 V45 V53 V58 V71 ? 11 110 112

1 1 1 1 1 1 1 1 0 0 0 0

141 150 151 154 155 156 157 158 160 161 162 164

0 0 0 0 0 0 0 0 0 0 0 0

171 174 180 182 183 184 185 187 188 189 191 193

0 0 0 0 0 0 0 0 0 0 0 0

196 197 199 (Other)

0 0 0 0

> summary(Maindataframe2$weight)

? [0-25) [100-125) [125-150) [150-175) [25-50) [50-75) [75-100)

0 2 28 7 1 4 34 57

> mytable <- table(Maindataframe2$weight)

> lbls <- paste(names(mytable), "\n", mytable, sep="")

> pie(mytable,labels = lbls, col=rainbow(length(lbls)),

+ main="Weight\n (After Cleaning)")

> mytable <- table(Maindataframe2$payer\_code)

> lbls <- paste(names(mytable), "\n", mytable, sep="")

> pie(mytable,labels = lbls, col=rainbow(length(lbls)),

+ main="Payer Code\n (After Cleaning)")

> summary(Maindataframe2$payer\_code)

? BC CH CM CP DM HM MC MD OG OT PO SI SP UN WC

0 8 0 3 11 7 5 72 1 0 0 0 0 0 26 0

>

>

>

>

> Maindataframe2$time\_in\_hospital<-as.numeric(Maindataframe2$time\_in\_hospital)

>

> Maindataframe2$number\_outpatient<-as.numeric(Maindataframe2$number\_outpatient)

>

> class(Maindataframe2$number\_outpatient)

[1] "numeric"

> Maindataframe2$number\_emergency<-as.numeric(Maindataframe2$number\_emergency)

>

> Maindataframe2$number\_inpatient<-as.numeric(Maindataframe2$number\_inpatient)

> unique(Maindataframe$race, incomparables = FALSE, fromLast = FALSE, nmax = NA)

[1] Caucasian AfricanAmerican Hispanic Other Asian ?

Levels: ? AfricanAmerican Asian Caucasian Hispanic Other

>

> tab1 <- xtabs(~race + readmitted, data = Maindataframe)

>

> barplot(tab1, main="Data Distribution by Readmitted patients vs Race results",xlab="Race Results", col=c("darkblue","red","blue","orange"),legend = rownames(tab1), beside=TRUE)

> install.packages("vcd")

Installing package into ‘C:/Users/Tanya Peddi/Documents/R/win-library/3.3’

(as ‘lib’ is unspecified)

trying URL 'https://cran.rstudio.com/bin/windows/contrib/3.3/vcd\_1.4-3.zip'

Content type 'application/zip' length 1266883 bytes (1.2 MB)

downloaded 1.2 MB

package ‘vcd’ successfully unpacked and MD5 sums checked

The downloaded binary packages are in

C:\Users\Tanya Peddi\AppData\Local\Temp\RtmpqAfBH2\downloaded\_packages

> library(vcd)

Loading required package: grid

Warning message:

package ‘vcd’ was built under R version 3.3.3

> summary(assocstats(tab1))

Call: xtabs(formula = ~race + readmitted, data = Maindataframe)

Number of cases in table: 10000

Number of factors: 2

Test for independence of all factors:

Chisq = 55.51, df = 5, p-value = 1.025e-10

X^2 df P(> X^2)

Likelihood Ratio 57.218 5 4.5592e-11

Pearson 55.511 5 1.0245e-10

Phi-Coefficient : NA

Contingency Coeff.: 0.074

Cramer's V : 0.075

> Maindataframe$admission\_type\_id[Maindataframe$admission\_type\_id == ""] <- "Not Available"

> unique(Maindataframe$admission\_type\_id, incomparables = FALSE, fromLast = FALSE, nmax = NA)

[1] Elective Urgent Not Available Emergency Not Mapped Newborn

Levels: Elective Emergency Newborn Not Available Not Mapped Urgent

> tab2 <- xtabs(~admission\_type\_id + readmitted, data = Maindataframe)

> barplot(tab2, main="Data Distribution by Readmitted patients vs Admission Type results",xlab="Admission Type Results", col=c("darkblue","red","blue","orange","green","yellow"),legend = rownames(tab2),args.legend = list(x ='topright', bty='n', inset=c(-0.25,0)), beside=TRUE)

>

>

>

> summary(assocstats(tab2))

Call: xtabs(formula = ~admission\_type\_id + readmitted, data = Maindataframe)

Number of cases in table: 10000

Number of factors: 2

Test for independence of all factors:

Chisq = NaN, df = 6, p-value = NA

Chi-squared approximation may be incorrect

X^2 df P(> X^2)

Likelihood Ratio 20.033 6 0.0027321

Pearson NaN 6 NaN

Phi-Coefficient : NA

Contingency Coeff.: NaN

Cramer's V : NaN

> unique(Maindataframe$admission\_source\_id, incomparables = FALSE, fromLast = FALSE, nmax = NA)

[1] Physician Referral

[3] Transfer from another health care facility Emergency Room

[5] Transfer from a Skilled Nursing Facility (SNF) Transfer from a hospital

[7] Not Mapped Clinic Referral

[9] HMO Referral Not Available

[11] Court/Law Enforcement

11 Levels: Clinic Referral Court/Law Enforcement Emergency Room HMO Referral Not Available Not Mapped ... Transfer from another health care facility

> Maindataframe$admission\_source\_id[Maindataframe$admission\_source\_id == ""] <- "Not Available"

> unique(Maindataframe$admission\_source\_id, incomparables = FALSE, fromLast = FALSE, nmax = NA)

[1] Physician Referral Not Available

[3] Transfer from another health care facility Emergency Room

[5] Transfer from a Skilled Nursing Facility (SNF) Transfer from a hospital

[7] Not Mapped Clinic Referral

[9] HMO Referral Court/Law Enforcement

11 Levels: Clinic Referral Court/Law Enforcement Emergency Room HMO Referral Not Available Not Mapped ... Transfer from another health care facility

> tab4 <- xtabs(~admission\_source\_id + readmitted, data = Maindataframe)

> summary(assocstats(tab4))

Call: xtabs(formula = ~admission\_source\_id + readmitted, data = Maindataframe)

Number of cases in table: 10000

Number of factors: 2

Test for independence of all factors:

Chisq = NaN, df = 10, p-value = NA

Chi-squared approximation may be incorrect

X^2 df P(> X^2)

Likelihood Ratio 116.91 10 0

Pearson NaN 10 NaN

Phi-Coefficient : NA

Contingency Coeff.: NaN

Cramer's V : NaN

> barplot(tab4, main="Data Distribution by Readmitted patients vs Admission Source results",xlab="Admission Source Results", col=c("darkblue","red","blue","orange","green","yellow"),legend = rownames(tab4),args.legend = list(x ='topright', bty='n', inset=c(-0.25,0)), beside=TRUE)

> unique(Maindataframe$medical\_specialty, incomparables = FALSE, fromLast = FALSE, nmax = NA)

[1] Surgery-Neuro ? Family/GeneralPractice

[4] Psychiatry Cardiology InternalMedicine

[7] Surgery-Cardiovascular/Thoracic Nephrology Emergency/Trauma

[10] Gastroenterology Orthopedics Cardiology-Pediatric

[13] PhysicalMedicineandRehabilitation Gynecology Pulmonology

[16] Surgery-General Pediatrics Orthopedics-Reconstructive

[19] Surgery-Pediatric Otolaryngology Pediatrics-CriticalCare

[22] Hematology/Oncology ObstetricsandGynecology Pediatrics-Endocrinology

[25] Surgery-Vascular Urology Neurology

[28] Radiologist Osteopath Surgery-Cardiovascular

[31] Psychology Oncology Endocrinology

[34] OutreachServices Podiatry Ophthalmology

[37] Hospitalist Radiology Obsterics&Gynecology-GynecologicOnco

[40] Surgery-Thoracic Surgeon Pathology

[43] Surgery-Plastic InfectiousDiseases Anesthesiology-Pediatric

[46] Pediatrics-Pulmonology Pediatrics-Hematology-Oncology Hematology

[49] Surgery-Colon&Rectal Surgery-PlasticwithinHeadandNeck Pediatrics-EmergencyMedicine

[52] Obstetrics PhysicianNotFound

53 Levels: ? Anesthesiology-Pediatric Cardiology Cardiology-Pediatric Emergency/Trauma ... Urology

> tab5 <- xtabs(~medical\_specialty + readmitted, data = Maindataframe)

> summary(assocstats(tab5))

Call: xtabs(formula = ~medical\_specialty + readmitted, data = Maindataframe)

Number of cases in table: 5900

Number of factors: 2

Test for independence of all factors:

Chisq = NaN, df = 52, p-value = NA

Chi-squared approximation may be incorrect

X^2 df P(> X^2)

Likelihood Ratio 177.77 52 1.2212e-15

Pearson NaN 52 NaN

Phi-Coefficient : NA

Contingency Coeff.: NaN

Cramer's V : NaN

> barplot(tab5, main="Data Distribution by Readmitted patients vs Medical Specialty results",xlab="Medical Specialty", col=c("darkblue","red","blue","orange","green","yellow"),legend = rownames(tab5),args.legend = list(x ='topright', bty='n', inset=c(-0.25,0)), beside=TRUE)

> numericaldata\_regression <- lm(readmitted ~ num\_lab\_procedures + num\_procedures+ num\_medications + number\_inpatient + number\_emergency + number\_outpatient + number\_diagnoses + time\_in\_hospital, data = Maindataframe)

> summary(numericaldata\_regression)

Call:

lm(formula = readmitted ~ num\_lab\_procedures + num\_procedures +

num\_medications + number\_inpatient + number\_emergency + number\_outpatient +

number\_diagnoses + time\_in\_hospital, data = Maindataframe)

Residuals:

Min 1Q Median 3Q Max

-1.2233 -0.3876 -0.2772 0.5553 0.8400

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 0.0936759 0.0192539 4.865 1.16e-06 \*\*\*

num\_lab\_procedures 0.0012808 0.0002580 4.965 6.99e-07 \*\*\*

num\_procedures -0.0098147 0.0030838 -3.183 0.00146 \*\*

num\_medications 0.0005334 0.0007073 0.754 0.45076

number\_inpatient 0.0929246 0.0057338 16.206 < 2e-16 \*\*\*

number\_emergency 0.0279140 0.0074492 3.747 0.00018 \*\*\*

number\_outpatient 0.0201957 0.0042870 4.711 2.50e-06 \*\*\*

number\_diagnoses 0.0300699 0.0024863 12.094 < 2e-16 \*\*\*

time\_in\_hospital -0.0006899 0.0018486 -0.373 0.70899

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4739 on 9991 degrees of freedom

Multiple R-squared: 0.06239, Adjusted R-squared: 0.06164

F-statistic: 83.1 on 8 and 9991 DF, p-value: < 2.2e-16

> unique(Maindataframe$max\_glu\_serum, incomparables = FALSE, fromLast = FALSE, nmax = NA)

[1] None >200 Norm >300

Levels: >200 >300 None Norm

> tab6 <- xtabs(~max\_glu\_serum + readmitted, data = Maindataframe)

> summary(assocstats(tab6))

Call: xtabs(formula = ~max\_glu\_serum + readmitted, data = Maindataframe)

Number of cases in table: 10000

Number of factors: 2

Test for independence of all factors:

Chisq = 3.828, df = 3, p-value = 0.2806

X^2 df P(> X^2)

Likelihood Ratio 3.8719 3 0.27563

Pearson 3.8283 3 0.28062

Phi-Coefficient : NA

Contingency Coeff.: 0.02

Cramer's V : 0.02

> barplot(tab6, main="Data Distribution by Readmitted patients vs Level of Glucose Serum results",xlab="Maximum Glucose Serum", col=c("darkblue","red","blue","orange"),legend = rownames(tab6),args.legend = list(x ='topright', bty='n', inset=c(-0.25,0)), beside=TRUE)

>

> unique(Maindataframe$A1Cresult, incomparables = FALSE, fromLast = FALSE, nmax = NA)

[1] None Norm >7 >8

Levels: >7 >8 None Norm

> tab7 <- xtabs(~A1Cresult + readmitted, data = Maindataframe)

> summary(assocstats(tab7))

Call: xtabs(formula = ~A1Cresult + readmitted, data = Maindataframe)

Number of cases in table: 10000

Number of factors: 2

Test for independence of all factors:

Chisq = 6.619, df = 3, p-value = 0.0851

X^2 df P(> X^2)

Likelihood Ratio 6.6816 3 0.082771

Pearson 6.6186 3 0.085101

Phi-Coefficient : NA

Contingency Coeff.: 0.026

Cramer's V : 0.026

> barplot(tab7, main="Data Distribution by Readmitted patients vs A1C test results",xlab="A1C Test results", col=c("darkblue","red","blue","orange"),legend = rownames(tab7),args.legend = list(x ='topright', bty='n', inset=c(-0.25,0)), beside=TRUE)

>

> unique(Maindataframe$metformin, incomparables = FALSE, fromLast = FALSE, nmax = NA)

[1] No Steady Down Up

Levels: Down No Steady Up

> tab8 <- xtabs(~metformin + readmitted, data = Maindataframe)

> summary(assocstats(tab8))

Call: xtabs(formula = ~metformin + readmitted, data = Maindataframe)

Number of cases in table: 10000

Number of factors: 2

Test for independence of all factors:

Chisq = 6.622, df = 3, p-value = 0.08497

X^2 df P(> X^2)

Likelihood Ratio 6.8176 3 0.077946

Pearson 6.6221 3 0.084971

Phi-Coefficient : NA

Contingency Coeff.: 0.026

Cramer's V : 0.026

>

> barplot(tab8, main="Data Distribution by Readmitted patients vs Metformin levels results",xlab="Metformin results", col=c("violet","red","pink","blue"),legend = rownames(tab8),args.legend = list(x ='topright', bty='n', inset=c(-0.25,0)), beside=TRUE)

>

>

> unique(Maindataframe$repaglinide, incomparables = FALSE, fromLast = FALSE, nmax = NA)

[1] No Steady Up Down

Levels: Down No Steady Up

> tab9 <- xtabs(~repaglinide + readmitted, data = Maindataframe)

> summary(assocstats(tab9))

Call: xtabs(formula = ~repaglinide + readmitted, data = Maindataframe)

Number of cases in table: 10000

Number of factors: 2

Test for independence of all factors:

Chisq = 14.505, df = 3, p-value = 0.002293

Chi-squared approximation may be incorrect

X^2 df P(> X^2)

Likelihood Ratio 14.261 3 0.0025708

Pearson 14.505 3 0.0022926

Phi-Coefficient : NA

Contingency Coeff.: 0.038

Cramer's V : 0.038

> barplot(tab9, main="Data Distribution by Readmitted patients vs Repaglinide levels results",xlab="Repaglinide results", col=c("violet","red","pink","blue"),legend = rownames(tab9),args.legend = list(x ='topright', bty='n', inset=c(-0.25,0)), beside=TRUE)

>

> unique(Maindataframe$nateglinide, incomparables = FALSE, fromLast = FALSE, nmax = NA)

[1] No Steady Down Up

Levels: Down No Steady Up

> tab10 <- xtabs(~nateglinide + readmitted, data = Maindataframe)

> summary(assocstats(tab10))

Call: xtabs(formula = ~nateglinide + readmitted, data = Maindataframe)

Number of cases in table: 10000

Number of factors: 2

Test for independence of all factors:

Chisq = 4.841, df = 3, p-value = 0.1839

Chi-squared approximation may be incorrect

X^2 df P(> X^2)

Likelihood Ratio 5.4606 3 0.14101

Pearson 4.8405 3 0.18385

Phi-Coefficient : NA

Contingency Coeff.: 0.022

Cramer's V : 0.022

> barplot(tab10, main="Data Distribution by Readmitted patients vs Nateglinide levels results",xlab="Nateglinide results", col=c("violet","yellow","pink","blue"),legend = rownames(tab10),args.legend = list(x ='topright', bty='n', inset=c(-0.25,0)), beside=TRUE)

>

> unique(Maindataframe$chlorpropamide, incomparables = FALSE, fromLast = FALSE, nmax = NA)

[1] No Steady Up

Levels: No Steady Up

> tab11 <- xtabs(~chlorpropamide + readmitted, data = Maindataframe)

> summary(assocstats(tab11))

Call: xtabs(formula = ~chlorpropamide + readmitted, data = Maindataframe)

Number of cases in table: 10000

Number of factors: 2

Test for independence of all factors:

Chisq = 1.5428, df = 2, p-value = 0.4624

Chi-squared approximation may be incorrect

X^2 df P(> X^2)

Likelihood Ratio 1.8707 2 0.39244

Pearson 1.5428 2 0.46237

Phi-Coefficient : NA

Contingency Coeff.: 0.012

Cramer's V : 0.012

> barplot(tab11, main="Data Distribution by Readmitted patients vs Chlorpropamide levels results",xlab="Chlorpropamide results", col=c("violet","green","pink"),legend = rownames(tab11),args.legend = list(x ='topright', bty='n', inset=c(-0.25,0)), beside=TRUE)

>

> unique(Maindataframe$glimepiride, incomparables = FALSE, fromLast = FALSE, nmax = NA)

[1] No Steady Up Down

Levels: Down No Steady Up

> tab12 <- xtabs(~glimepiride + readmitted, data = Maindataframe)

> summary(assocstats(tab12))

Call: xtabs(formula = ~glimepiride + readmitted, data = Maindataframe)

Number of cases in table: 10000

Number of factors: 2

Test for independence of all factors:

Chisq = 4.521, df = 3, p-value = 0.2104

Chi-squared approximation may be incorrect

X^2 df P(> X^2)

Likelihood Ratio 4.4471 3 0.21706

Pearson 4.5209 3 0.21043

Phi-Coefficient : NA

Contingency Coeff.: 0.021

Cramer's V : 0.021

> barplot(tab12, main="Data Distribution by Readmitted patients vs Glimepiride levels results",xlab="Glimepiride results", col=c("violet","green","pink","red"),legend = rownames(tab12),args.legend = list(x ='topright', bty='n', inset=c(-0.25,0)), beside=TRUE)

>

>

> unique(Maindataframe$acetohexamide, incomparables = FALSE, fromLast = FALSE, nmax = NA)

[1] No

Levels: No

> unique(Maindataframe$glipizide, incomparables = FALSE, fromLast = FALSE, nmax = NA)

[1] No Steady Up Down

Levels: Down No Steady Up

> tab13 <- xtabs(~glipizide + readmitted, data = Maindataframe)

> summary(assocstats(tab13))

Call: xtabs(formula = ~glipizide + readmitted, data = Maindataframe)

Number of cases in table: 10000

Number of factors: 2

Test for independence of all factors:

Chisq = 5.524, df = 3, p-value = 0.1372

X^2 df P(> X^2)

Likelihood Ratio 5.5067 3 0.13824

Pearson 5.5241 3 0.13720

Phi-Coefficient : NA

Contingency Coeff.: 0.023

Cramer's V : 0.024

> barplot(tab13, main="Data Distribution by Readmitted patients vs Glipizide levels results",xlab="Glipizide results", col=c("violet","green","pink","red"),legend = rownames(tab13),args.legend = list(x ='topright', bty='n', inset=c(-0.25,0)), beside=TRUE)

>

>

> unique(Maindataframe$glyburide, incomparables = FALSE, fromLast = FALSE, nmax = NA)

[1] No Steady Down Up

Levels: Down No Steady Up

> tab14 <- xtabs(~glyburide + readmitted, data = Maindataframe)

> summary(assocstats(tab14))

Call: xtabs(formula = ~glyburide + readmitted, data = Maindataframe)

Number of cases in table: 10000

Number of factors: 2

Test for independence of all factors:

Chisq = 3.716, df = 3, p-value = 0.2938

X^2 df P(> X^2)

Likelihood Ratio 3.6503 3 0.30179

Pearson 3.7159 3 0.29382

Phi-Coefficient : NA

Contingency Coeff.: 0.019

Cramer's V : 0.019

> barplot(tab14, main="Data Distribution by Readmitted patients vs Glyburide levels results",xlab="Glyburide results", col=c("violet","pink","green","red"),legend = rownames(tab14),args.legend = list(x ='topright', bty='n', inset=c(-0.25,0)), beside=TRUE)

>

> unique(Maindataframe$tolbutamide, incomparables = FALSE, fromLast = FALSE, nmax = NA)

[1] No Steady

Levels: No Steady

> tab15 <- xtabs(~tolbutamide + readmitted, data = Maindataframe)

> summary(assocstats(tab15))

Call: xtabs(formula = ~tolbutamide + readmitted, data = Maindataframe)

Number of cases in table: 10000

Number of factors: 2

Test for independence of all factors:

Chisq = 0.05004, df = 1, p-value = 0.823

Chi-squared approximation may be incorrect

X^2 df P(> X^2)

Likelihood Ratio 0.051125 1 0.82112

Pearson 0.050039 1 0.82300

Phi-Coefficient : 0.002

Contingency Coeff.: 0.002

Cramer's V : 0.002

> barplot(tab15, main="Data Distribution by Readmitted patients vs Tolbutamide levels results",xlab="Tolbutamide results", col=c("violet","red"),legend = rownames(tab15),args.legend = list(x ='topright', bty='n', inset=c(-0.25,0)), beside=TRUE)

>

> unique(Maindataframe$pioglitazone, incomparables = FALSE, fromLast = FALSE, nmax = NA)

[1] No Steady Up Down

Levels: Down No Steady Up

> tab16 <- xtabs(~pioglitazone + readmitted, data = Maindataframe)

> barplot(tab16, main="Data Distribution by Readmitted patients vs Pioglitazone levels results",xlab="Pioglitazone results", col=c("violet","red","green","blue"),legend = rownames(tab16),args.legend = list(x ='topright', bty='n', inset=c(-0.25,0)), beside=TRUE)

> summary(assocstats(tab16))

Call: xtabs(formula = ~pioglitazone + readmitted, data = Maindataframe)

Number of cases in table: 10000

Number of factors: 2

Test for independence of all factors:

Chisq = 30.993, df = 3, p-value = 8.529e-07

X^2 df P(> X^2)

Likelihood Ratio 30.401 3 1.1364e-06

Pearson 30.993 3 8.5293e-07

Phi-Coefficient : NA

Contingency Coeff.: 0.056

Cramer's V : 0.056

> unique(Maindataframe$rosiglitazone, incomparables = FALSE, fromLast = FALSE, nmax = NA)

[1] No Steady Up Down

Levels: Down No Steady Up

> tab17 <- xtabs(~rosiglitazone + readmitted, data = Maindataframe)

> summary(assocstats(tab17))

Call: xtabs(formula = ~rosiglitazone + readmitted, data = Maindataframe)

Number of cases in table: 10000

Number of factors: 2

Test for independence of all factors:

Chisq = 2.3247, df = 3, p-value = 0.5078

Chi-squared approximation may be incorrect

X^2 df P(> X^2)

Likelihood Ratio 2.3300 3 0.5068

Pearson 2.3247 3 0.5078

Phi-Coefficient : NA

Contingency Coeff.: 0.015

Cramer's V : 0.015

> barplot(tab17, main="Data Distribution by Readmitted patients vs Rosiglitazone levels results",xlab="Rosiglitazone results", col=c("violet","red","green","blue"),legend = rownames(tab17),args.legend = list(x ='topright', bty='n', inset=c(-0.25,0)), beside=TRUE)

>

> unique(Maindataframe$acarbose, incomparables = FALSE, fromLast = FALSE, nmax = NA)

[1] No Steady Up

Levels: No Steady Up

> tab18 <- xtabs(~acarbose + readmitted, data = Maindataframe)

> summary(assocstats(tab18))

Call: xtabs(formula = ~acarbose + readmitted, data = Maindataframe)

Number of cases in table: 10000

Number of factors: 2

Test for independence of all factors:

Chisq = 0.6687, df = 2, p-value = 0.7158

Chi-squared approximation may be incorrect

X^2 df P(> X^2)

Likelihood Ratio 1.02170 2 0.59998

Pearson 0.66865 2 0.71582

Phi-Coefficient : NA

Contingency Coeff.: 0.008

Cramer's V : 0.008

> barplot(tab18, main="Data Distribution by Readmitted patients vs Acarbose levels results",xlab="Acarbose results", col=c("violet","red","green"),legend = rownames(tab18),args.legend = list(x ='topright', bty='n', inset=c(-0.25,0)), beside=TRUE)

>

> unique(Maindataframe$miglitol, incomparables = FALSE, fromLast = FALSE, nmax = NA)

[1] No Steady Down Up

Levels: Down No Steady Up

> tab19 <- xtabs(~miglitol + readmitted, data = Maindataframe)

> summary(assocstats(tab19))

Call: xtabs(formula = ~miglitol + readmitted, data = Maindataframe)

Number of cases in table: 10000

Number of factors: 2

Test for independence of all factors:

Chisq = 3.961, df = 3, p-value = 0.2657

Chi-squared approximation may be incorrect

X^2 df P(> X^2)

Likelihood Ratio 4.5933 3 0.20412

Pearson 3.9609 3 0.26571

Phi-Coefficient : NA

Contingency Coeff.: 0.02

Cramer's V : 0.02

> barplot(tab19, main="Data Distribution by Readmitted patients vs Miglitol levels results",xlab="Miglitol results", col=c("violet","magenta","green","yellow"),legend = rownames(tab19),args.legend = list(x ='topright', bty='n', inset=c(-0.25,0)), beside=TRUE)

>

> unique(Maindataframe$troglitazone, incomparables = FALSE, fromLast = FALSE, nmax = NA)

[1] No

Levels: No

> unique(Maindataframe$tolazamide, incomparables = FALSE, fromLast = FALSE, nmax = NA)

[1] No Steady

Levels: No Steady

> tab20 <- xtabs(~tolazamide + readmitted, data = Maindataframe)

> summary(assocstats(tab20))

Call: xtabs(formula = ~tolazamide + readmitted, data = Maindataframe)

Number of cases in table: 10000

Number of factors: 2

Test for independence of all factors:

Chisq = 2.6291, df = 1, p-value = 0.1049

Chi-squared approximation may be incorrect

X^2 df P(> X^2)

Likelihood Ratio 4.0411 1 0.044404

Pearson 2.6291 1 0.104924

Phi-Coefficient : 0.016

Contingency Coeff.: 0.016

Cramer's V : 0.016

> barplot(tab20, main="Data Distribution by Readmitted patients vs Tolazamide levels results",xlab="Tolazamide results", col=c("magenta","green"),legend = rownames(tab20),args.legend = list(x ='topright', bty='n', inset=c(-0.25,0)), beside=TRUE)

>

> unique(Maindataframe$examide, incomparables = FALSE, fromLast = FALSE, nmax = NA)

[1] No

Levels: No

> unique(Maindataframe$citoglipton, incomparables = FALSE, fromLast = FALSE, nmax = NA)

[1] No

Levels: No

> unique(Maindataframe$insulin, incomparables = FALSE, fromLast = FALSE, nmax = NA)

[1] No Steady Down Up

Levels: Down No Steady Up

> tab21 <- xtabs(~insulin + readmitted, data = Maindataframe)

> summary(assocstats(tab21))

Call: xtabs(formula = ~insulin + readmitted, data = Maindataframe)

Number of cases in table: 10000

Number of factors: 2

Test for independence of all factors:

Chisq = 31.67, df = 3, p-value = 6.136e-07

X^2 df P(> X^2)

Likelihood Ratio 31.251 3 7.5263e-07

Pearson 31.672 3 6.1363e-07

Phi-Coefficient : NA

Contingency Coeff.: 0.056

Cramer's V : 0.056

> barplot(tab21, main="Data Distribution by Readmitted patients vs Insulin levels results",xlab="Insulin results", col=c("yellow","magenta","green","blue"),legend = rownames(tab21),args.legend = list(x ='topright', bty='n', inset=c(-0.25,0)), beside=TRUE)

>

> unique(Maindataframe$glyburide.metformin, incomparables = FALSE, fromLast = FALSE, nmax = NA)

[1] No Steady Down Up

Levels: Down No Steady Up

> tab22 <- xtabs(~glyburide.metformin + readmitted, data = Maindataframe)

> summary(assocstats(tab22))

Call: xtabs(formula = ~glyburide.metformin + readmitted, data = Maindataframe)

Number of cases in table: 10000

Number of factors: 2

Test for independence of all factors:

Chisq = 2.6752, df = 3, p-value = 0.4445

Chi-squared approximation may be incorrect

X^2 df P(> X^2)

Likelihood Ratio 3.7245 3 0.29279

Pearson 2.6752 3 0.44446

Phi-Coefficient : NA

Contingency Coeff.: 0.016

Cramer's V : 0.016

> barplot(tab22, main="Data Distribution by Readmitted patients vs Glyburide Metformin levels results",xlab="Glyburide Metformin results", col=c("yellow","magenta","green","blue"),legend = rownames(tab22),args.legend = list(x ='topright', bty='n', inset=c(-0.25,0)), beside=TRUE)

>

> unique(Maindataframe$glipizide.metformin, incomparables = FALSE, fromLast = FALSE, nmax = NA)

[1] No Steady

Levels: No Steady

> tab23 <- xtabs(~glipizide.metformin + readmitted, data = Maindataframe)

> summary(assocstats(tab23))

Call: xtabs(formula = ~glipizide.metformin + readmitted, data = Maindataframe)

Number of cases in table: 10000

Number of factors: 2

Test for independence of all factors:

Chisq = 0.08955, df = 1, p-value = 0.7647

Chi-squared approximation may be incorrect

X^2 df P(> X^2)

Likelihood Ratio 0.087606 1 0.76724

Pearson 0.089552 1 0.76475

Phi-Coefficient : 0.003

Contingency Coeff.: 0.003

Cramer's V : 0.003

> barplot(tab23, main="Data Distribution by Readmitted patients vs Glipizide Metformin levels results",xlab="Glipizide Metformin results", col=c("yellow","magenta"),legend = rownames(tab23),args.legend = list(x ='topright', bty='n', inset=c(-0.25,0)), beside=TRUE)

>

> unique(Maindataframe$glimepiride.pioglitazone, incomparables = FALSE, fromLast = FALSE, nmax = NA)

[1] No

Levels: No

> unique(Maindataframe$metformin.rosiglitazone, incomparables = FALSE, fromLast = FALSE, nmax = NA)

[1] No

Levels: No

> unique(Maindataframe$metformin.pioglitazone, incomparables = FALSE, fromLast = FALSE, nmax = NA)

[1] No

Levels: No

> unique(Maindataframe$change, incomparables = FALSE, fromLast = FALSE, nmax = NA)

[1] No Ch

Levels: Ch No

> unique(Maindataframe$diabetesMed, incomparables = FALSE, fromLast = FALSE, nmax = NA)

[1] No Yes

Levels: No Yes

> tab24 <- xtabs(~diabetesMed + readmitted, data = Maindataframe)

> summary(assocstats(tab24))

Call: xtabs(formula = ~diabetesMed + readmitted, data = Maindataframe)

Number of cases in table: 10000

Number of factors: 2

Test for independence of all factors:

Chisq = 28.281, df = 1, p-value = 1.049e-07

X^2 df P(> X^2)

Likelihood Ratio 28.561 1 9.0785e-08

Pearson 28.281 1 1.0491e-07

Phi-Coefficient : 0.053

Contingency Coeff.: 0.053

Cramer's V : 0.053

> barplot(tab24, main="Data Distribution by Readmitted patients vs Diabetes Med Status results",xlab="Diabetes Med Status", col=c("yellow","magenta"),legend = rownames(tab24),args.legend = list(x ='topright', bty='n', inset=c(-0.25,0)), beside=TRUE)

>

> tab25 <- xtabs(~gender + readmitted, data = Maindataframe)

> summary(assocstats(tab25))

Call: xtabs(formula = ~gender + readmitted, data = Maindataframe)

Number of cases in table: 10000

Number of factors: 2

Test for independence of all factors:

Chisq = 1.5848, df = 1, p-value = 0.2081

X^2 df P(> X^2)

Likelihood Ratio 1.5853 1 0.20800

Pearson 1.5848 1 0.20807

Phi-Coefficient : 0.013

Contingency Coeff.: 0.013

Cramer's V : 0.013

> barplot(tab24, main="Data Distribution by Readmitted patients vs Gender results",xlab="Gender", col=c("yellow","magenta"),legend = rownames(tab25),args.legend = list(x ='topright', bty='n', inset=c(-0.25,0)), beside=TRUE)

>

NOTE: Maindataframe3 consists of only those that have gotten a high chi square value - race, admitted\_type\_id,medical\_speciality, num\_lab\_procedures, num\_procedures,num\_diagnoses, repagl..,piopa...., insulin, diabetesMed

> trial$readmitted <- ifelse(trial$readmitted=="FALSE",0,1)

> View(trial)

> res3 <- hc(trial)

> plot(res3)

.

.

.

.

.

> trial10 <- Maindataframe[c(2,3,6,7,8,11,22,23,24,25,26,27,28,29,30,31,32,33,34,35,36,37,38,39,40,41,42,43,44,45,46,47,48,49)]

> trial10$readmitted <- ifelse(trial10$readmitted=="FALSE",0,1)

> res12 <- hc(trial10)

Error in check.data(x) : the data set contains NULL/NaN/NA values.

> trial10 <- na.omit(trial10)

> res12 <- hc(trial10)

> plot(res12)

> resmain <- hc(trial)

> plot(resmain)

Error in plot.new() : figure margins too large

> plot(resmain)

> fittedbn <- bn.fit(resmain, data = trial)

> fittedbn

> trial10 <- Maindataframe[c(22,23,24,25,28,30,31,33,34,35,41,43,47,48,49)]

> trial10$readmitted <- ifelse(trial10$readmitted=="FALSE",0,1)

> View(Maindataframe2)

> View(trial10)

> res10 <- hc(trial10)

Error: could not find function "hc"

> library(bnlearn)

Attaching package: ‘bnlearn’

The following object is masked from ‘package:stats’:

sigma

Warning message:

package ‘bnlearn’ was built under R version 3.3.3

> res10 <- hc(trial10)

> plot(res10)

> fittedbn <- bn.fit(res10, data = trial10)

> fittedbn

Bayesian network parameters

Parameters of node max\_glu\_serum (multinomial distribution)

Conditional probability table:

A1Cresult

max\_glu\_serum >7 >8 None Norm

>200 0.005665722 0.003658537 0.022675737 0.004464286

>300 0.008498584 0.010975610 0.013605442 0.000000000

None 0.977337110 0.980487805 0.924215300 0.988839286

Norm 0.008498584 0.004878049 0.039503521 0.006696429

Parameters of node A1Cresult (multinomial distribution)

Conditional probability table:

diabetesMed

A1Cresult No Yes

>7 0.03172086 0.03650709

>8 0.03449643 0.09802086

None 0.87390960 0.82575555

Norm 0.05987312 0.03971650

Parameters of node metformin (multinomial distribution)

Conditional probability table:

Down No Steady Up

0.0051 0.8011 0.1816 0.0122

Parameters of node repaglinide (multinomial distribution)

Conditional probability table:

change

repaglinide Ch No

Down 0.001169317 0.000000000

No 0.975678204 0.995457722

Steady 0.020112254 0.004542278

Up 0.003040225 0.000000000

Parameters of node glimepiride (multinomial distribution)

Conditional probability table:

, , diabetesMed = No

change

glimepiride Ch No

Down 0.000000000

No 1.000000000

Steady 0.000000000

Up 0.000000000

, , diabetesMed = Yes

change

glimepiride Ch No

Down 0.002104771 0.000000000

No 0.916744621 0.957838851

Steady 0.073666978 0.042161149

Up 0.007483630 0.000000000

Parameters of node glipizide (multinomial distribution)

Conditional probability table:

, , diabetesMed = No

insulin

glipizide Down No Steady Up

Down 0.000000000

No 1.000000000

Steady 0.000000000

Up 0.000000000

, , diabetesMed = Yes

insulin

glipizide Down No Steady Up

Down 0.004246285 0.007963595 0.007523716 0.004750594

No 0.894904459 0.729996208 0.864245993 0.907363420

Steady 0.088110403 0.244975351 0.118089630 0.083135392

Up 0.012738854 0.017064846 0.010140661 0.004750594

Parameters of node glyburide (multinomial distribution)

Conditional probability table:

, , change = Ch, diabetesMed = No

insulin

glyburide Down No Steady Up

Down

No

Steady

Up

, , change = No, diabetesMed = No

insulin

glyburide Down No Steady Up

Down 0.000000000

No 1.000000000

Steady 0.000000000

Up 0.000000000

, , change = Ch, diabetesMed = Yes

insulin

glyburide Down No Steady Up

Down 0.003184713 0.035339064 0.011764706 0.004750594

No 0.944798301 0.605539637 0.784083045 0.939429929

Steady 0.047770701 0.307545368 0.184083045 0.049881235

Up 0.004246285 0.051575931 0.020069204 0.005938242

, , change = No, diabetesMed = Yes

insulin

glyburide Down No Steady Up

Down 0.000000000 0.000000000

No 0.774842767 1.000000000

Steady 0.225157233 0.000000000

Up 0.000000000 0.000000000

Parameters of node pioglitazone (multinomial distribution)

Conditional probability table:

, , diabetesMed = No

change

pioglitazone Ch No

Down 0.000000000

No 1.000000000

Steady 0.000000000

Up 0.000000000

, , diabetesMed = Yes

change

pioglitazone Ch No

Down 0.003040225 0.000000000

No 0.866463985 0.957838851

Steady 0.123713751 0.042161149

Up 0.006782039 0.000000000

Parameters of node rosiglitazone (multinomial distribution)

Conditional probability table:

, , diabetesMed = No

change

rosiglitazone Ch No

Down 0.000000000

No 1.000000000

Steady 0.000000000

Up 0.000000000

, , diabetesMed = Yes

change

rosiglitazone Ch No

Down 0.002104771 0.000000000

No 0.850093545 0.962523423

Steady 0.142656688 0.037476577

Up 0.005144995 0.000000000

Parameters of node acarbose (multinomial distribution)

Conditional probability table:

change

acarbose Ch No

No 0.9932179607 0.9994758910

Steady 0.0065481759 0.0005241090

Up 0.0002338634 0.0000000000

Parameters of node insulin (multinomial distribution)

Conditional probability table:

Down No Steady Up

0.0942 0.5159 0.3057 0.0842

Parameters of node glipizide.metformin (multinomial distribution)

Conditional probability table:

No Steady

0.9998 0.0002

Parameters of node change (multinomial distribution)

Conditional probability table:

, , glipizide = Down, insulin = Down

metformin

change Down No Steady Up

Ch 1.00000000 1.00000000 1.00000000

No 0.00000000 0.00000000 0.00000000

, , glipizide = No, insulin = Down

metformin

change Down No Steady Up

Ch 1.00000000 1.00000000 1.00000000 1.00000000

No 0.00000000 0.00000000 0.00000000 0.00000000

, , glipizide = Steady, insulin = Down

metformin

change Down No Steady Up

Ch 1.00000000 1.00000000 1.00000000 1.00000000

No 0.00000000 0.00000000 0.00000000 0.00000000

, , glipizide = Up, insulin = Down

metformin

change Down No Steady Up

Ch 1.00000000 1.00000000

No 0.00000000 0.00000000

, , glipizide = Down, insulin = No

metformin

change Down No Steady Up

Ch 1.00000000 1.00000000

No 0.00000000 0.00000000

, , glipizide = No, insulin = No

metformin

change Down No Steady Up

Ch 1.00000000 0.05860806 0.51266586 1.00000000

No 0.00000000 0.94139194 0.48733414 0.00000000

, , glipizide = Steady, insulin = No

metformin

change Down No Steady Up

Ch 1.00000000 0.23700624 1.00000000 1.00000000

No 0.00000000 0.76299376 0.00000000 0.00000000

, , glipizide = Up, insulin = No

metformin

change Down No Steady Up

Ch 1.00000000 1.00000000 1.00000000

No 0.00000000 0.00000000 0.00000000

, , glipizide = Down, insulin = Steady

metformin

change Down No Steady Up

Ch 1.00000000 1.00000000 1.00000000 1.00000000

No 0.00000000 0.00000000 0.00000000 0.00000000

, , glipizide = No, insulin = Steady

metformin

change Down No Steady Up

Ch 1.00000000 0.25023256 1.00000000 1.00000000

No 0.00000000 0.74976744 0.00000000 0.00000000

, , glipizide = Steady, insulin = Steady

metformin

change Down No Steady Up

Ch 1.00000000 1.00000000 1.00000000 1.00000000

No 0.00000000 0.00000000 0.00000000 0.00000000

, , glipizide = Up, insulin = Steady

metformin

change Down No Steady Up

Ch 1.00000000 1.00000000

No 0.00000000 0.00000000

, , glipizide = Down, insulin = Up

metformin

change Down No Steady Up

Ch 1.00000000

No 0.00000000

, , glipizide = No, insulin = Up

metformin

change Down No Steady Up

Ch 1.00000000 1.00000000 1.00000000 1.00000000

No 0.00000000 0.00000000 0.00000000 0.00000000

, , glipizide = Steady, insulin = Up

metformin

change Down No Steady Up

Ch 1.00000000 1.00000000

No 0.00000000 0.00000000

, , glipizide = Up, insulin = Up

metformin

change Down No Steady Up

Ch 1.00000000 1.00000000

No 0.00000000 0.00000000

Parameters of node diabetesMed (multinomial distribution)

Conditional probability table:

, , insulin = Down

metformin

diabetesMed Down No Steady Up

No 0.0000000 0.0000000 0.0000000 0.0000000

Yes 1.0000000 1.0000000 1.0000000 1.0000000

, , insulin = No

metformin

diabetesMed Down No Steady Up

No 0.0000000 0.6178344 0.0000000 0.0000000

Yes 1.0000000 0.3821656 1.0000000 1.0000000

, , insulin = Steady

metformin

diabetesMed Down No Steady Up

No 0.0000000 0.0000000 0.0000000 0.0000000

Yes 1.0000000 1.0000000 1.0000000 1.0000000

, , insulin = Up

metformin

diabetesMed Down No Steady Up

No 0.0000000 0.0000000 0.0000000 0.0000000

Yes 1.0000000 1.0000000 1.0000000 1.0000000

Parameters of node readmitted (conditional Gaussian distribution)

Conditional density: readmitted | diabetesMed

Coefficients:

0 1

(Intercept) 0.3517050 0.4116074

Standard deviation of the residuals:

0 1

0.4775971 0.4921576

Discrete parents' configurations:

diabetesMed

0 No

1. Yes