**M S Ramaiah Institute of Technology**

(An Autonomous Institute, Affiliated to VTU)

MSR Nagar, MSRIT post, Bangalore-54

A Dissertation Report on

Data Analytics on diabetic data set to predict and to take preventive measures

Submitted by

H.Rameshwari - 1MS12CS036  
 Nikhila Somu - 1MS12CS066  
 Chandrakala - 1MS12CS020  
Deeksha - 1MS12CS025

*in partial fulfillment for the award of the degree of*

# *Bachelor of Engineering in Computer Science & Engineering*



**DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING**

**M.S.RAMAIAH INSTITUTE OF TECHNOLOGY**

**(Autonomous Institute, Affiliated to VTU)**

**BANGALORE-560054**

[www.msrit.edu](http://www.msrit.edu), **May 2015**

# Abstract

Management of hyperglycemia in hospitalized patients has a significant bearing on outcome, in terms of both morbidity and mortality. However, there are few national assessments of diabetes care during hospitalization which could serve as a baseline for change. This analysis of a large clinical database (74 million unique encounters corresponding to 17 million unique patients) was undertaken to provide such an assessment and to find future directions which might lead to improvements in patient safety. Almost 70,000 inpatient diabetes encounters were identified with sufficient detail for analysis. Multivariable logistic regression was used to fit the relationship between the measurement of HbA1c and early readmission while controlling for covariates such as demographics, severity and type of the disease, and type of admission. Results show that the measurement of HbA1c was performed infrequently (18.4%) in the inpatient setting. The statistical model suggests that the relationship between the probability of readmission and the HbA1c measurement depends on the primary diagnosis. The data suggest further that the greater attention to diabetes reflected in HbA1c determination may improve patient outcomes and lower cost of inpatient care.

#### Contents

***Declaration i***

***Acknowledgements ii***

***Abstract***

1 **INTRODUCTION**

* 1. General Introduction
  2. Statement of the Problem
  3. Objectives of the project
  4. Project deliverables
  5. Current Scope
  6. Future Scope

1. **PROJECT ORGANIZATION**
   1. Software Process Models
   2. Roles and Responsibilities
2. **LITERATURE SURVEY**

3.1 Introduction

3.2 Main Body

3.3 Conclusion of Survey

1. **SOFTWARE REQUIREMENT SPECIFICATIONS**

4.1 Product Overview

4.2 External Interface Requirements

4.2.1 User Interfaces

4.2.2 Hardware Interfaces

4.2.3 Software Interfaces

4.2.4 Communication Interfaces

4.3 Functional Requirements

4.3.1 Functional Requirement 1.1

:

4.3.n Functional Requirement 1.n

4.4 Software System Attributes

4.4.1 Reliability

4.4.2 Availability

4.4.3 Security

4.4.4 Portability

4.4.5 Maintainability

4.4.6 Performance

4.5 Performance Requirements

4.6 Database Requirement

4.7 Design Constraints

4.8 Other Requirements

1. **DESIGN**
   1. Introduction
   2. Architecture Design
   3. Graphical User Interface
   4. Class Diagram and Classes (represent Inheritance, Aggregation and Association)
   5. Sequence Diagram
   6. Data flow diagram
   7. Metric calculation
2. **IMPLEMENTATION**
   1. Tools Introduction
   2. Technology Introduction
   3. Overall view of the project in terms of implementation
   4. Explanation of Algorithm and how it is been implemented
   5. Information about the implementation of Modules
3. **TESTING**
   1. Results and Snapshots
4. **CONCLUSION & SCOPE FOR FUTURE WORK**

1. **REFERENCES**

1. **INTRODUCTION**

**1.1 General Introduction**

It is increasingly recognized that the management of hyperglycemia in the hospitalized patient has a significant bearing on outcome, in terms of both morbidity and mortality . This recognition has led to the development of formalized protocols in the intensive care unit (ICU) setting with rigorous glucose targets in many institutions . However, the same cannot be said for most non-ICU inpatient admissions. Rather, anecdotal evidence suggests that inpatient management is arbitrary and often leads to either no treatment at all or wide fluctuations in glucose when traditional management strategies are employed. Although data are few, recent controlled trials have demonstrated that protocol-driven inpatient strategies can be both effective and safe [[4](http://www.hindawi.com/journals/bmri/2014/781670/#B4), [5](http://www.hindawi.com/journals/bmri/2014/781670/#B5)]. As such, implementation of protocols in the hospital setting is now recommended [[6](http://www.hindawi.com/journals/bmri/2014/781670/#B6), [7](http://www.hindawi.com/journals/bmri/2014/781670/#B7)]. However, there are few national assessments of diabetes care in the hospitalized patient which could serve as a baseline for change. 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. In particular, we examined the use of HbA1c as a marker of attention to diabetes care in a large number of individuals identified as having a diagnosis of diabetes mellitus. We hypothesize that measurement of HbA1c is associated with a reduction in readmission rates in individuals admitted to the hospital.

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. [[8](http://www.hindawi.com/journals/bmri/2014/781670/#B8)]. Additionally, analyzing external data is more challenging than analysis of results of a carefully designed experiment or trial, because one has no impact on how and what type of information was collected. Nonetheless, it is important to utilize these huge amounts of data to find new information/knowledge that is possibly not available anywhere.

**1.2 Statement of the Problem**

Data Analytics on diabetic data set to predict and to take preventive measures

**1.3 Scope and Future Work**

The future involves the in analyzing the dataset by using the advanced algorithm. It reduces the variations in the prediction and makes the performance faster. It includes removing the deviation, constructing better models with minimum possible deviation in the realistic and predicted data. We recognize that the results from the present analysis represent a preliminary observation with limitations intrinsic to such large health records which in future can be eliminated. In future, more number of district comparison may takes place and analysis may be done. The future work will reduce the redundant data, improves the performance, effectiveness and accuracy of the search. This extension will give the better result for the performance ratio.

1. **PROJECT ORGANIZATION**
   1. **Software Process Models**

Classification Modeling with Cost Sensitive Analysis Even though the accuracy is more than 90%, we considered in focusing of misclassification cost. The wrong prediction cases are still important criteria expected in medical diagnostics situation. Since in the case of diabetes, the cost of false negatives is higher than false positive, as the disease can progress very rapidly when the patient was left untreated. If the diabetes patients are diagnosed as non-diabetes patients, they will be untreated for years resulting being bad health and may lose their lives. It is much more serious than the non-diabetes patients are diagnosed as diabetes patients. Therefore, asymmetric classification cost was considered in this paper. The corrected prediction cases are fine. However, the patient who could delay in the correct treatment must have some cost penalty. The cost matrix given with this experiment. The asymmetric classification costs ratio of FN over those of FP at 5:1, 10:1, 15:1, 20:1, 25:1, and 30:1. The evaluation of classification models used these cost ratios including 1:1 ratio (symmetric cost). It demonstrates that classification model accuracy decreases when asymmetric cost ratio of FN: FP increases. These can be alternative choices based on certain situations. However, an appropriate cost sensitive ratio depends on discretion of doctors.

**3. LITERATURE SURVEY**

* 1. **Introduction**

India has a high prevalence of diabetes mellitus and the numbers are increasing at an alarming rate. In India alone, diabetes is expected to increase from 40.6 million in 2006 to 79.4 million by 2030 and the projected estimate of the people with diabetes worldwide is 354 million. This statistics clearly indicates that, out of 4 diabetic people in the world, one will be Indian. Other studies have shown that the prevalence of diabetes in urban Indian adults is about 12% and the Type 2 Diabetes is 4-6 time higher in urban than in rural areas. This growth in the urban areas is because of the increase in the rates of obesity which have tripled in the last two decades due to the change in life-style and lack of physical activity. Type 2 Diabetes (T2D) is strongly associated with morbidity and mortality and carries a heavy financial burden.

**3.2 Conclusion of Survey**

Diabetes mellitus and its complications have proven to be parts of the major public health problems in Finland and it has called for the need to carry out research works on how to minimize these problems. The main purpose of this thesis is to produce information about prevention of diabetes mellitus complications among adults and the aim is to produce a guide for the public use in terveysnetti pages. Systematic literature review was used to analyze the data. In this research, thirty recent journals were reviewed and five books were included. Academic databases such as Ensco, Science direct, OVID and Cinahi were used as sources of data. The study was based on the adults but the beneficiaries of the result could be any class. In the end, it was realized that efforts should be continuously made, in preventing the complications, rather than focusing on the treatment of the complications. Health information about how to prevent diabetes mellitus complication is given and the guide is developed separately. Possible researchable topics in the field of diabetes mellitus are also suggested as research areas for the future health students.

1. **SOFTWARE REQUIREMENT SPECIFICATIONS**

**4.1 External Interface Requirements**

**4.1.1 User Interface**

The system’s UI is designed to be spontaneous and user-friendly as possible. An error messages will prompt the user in case of an inaccuracy in input. Notifications through warning messages will serve to keep the user posted.

**4.1.2Hardware Interface**

* Ubuntu 14.04
* Enough memory
* Enough processing capabilities

**4.1.3Software Interface**

* PHP / R programming
* R Studio
* R packages

**4.1.4Communication Interface**

No additional specific communication interfaces are needed during the operation.

**4.2 Functional Requirements**

a. Analysis based on age group of patients, gender, medication, etc.

b. Chi squared testing on different variables.

c. Linear Regression.

d. Logistic regression

e. k-means clustering algorithm

**4.3. Software System Attributes**

**4.3.1 Reliability**: Software shall never crash or freeze. The unit shall inform the user to re-set the system if major software mal-function has been encounter by the system.

**4.3.2 Availability**: Data shall be always available when queried.

* + 1. **Security**: Data shall be only retrieved but not modified.
    2. **Portability:** No portability identified at this movement.
    3. **Maintainability**: The system shall be flexible enough to add new modules and upgrade the existing modules.
    4. **Performance**: queries should be executed in time.

**4.5 Performance Requirements**

* The response time should be less than 1 second
* Capability to handle 1 lakh+ data points.

**4.6 Database Requirement**

R is capable of storing and provides quick random access to huge amounts of structured data

**4.7 Design Constraints**

Additional hardware (memory) and software is required in future to accommodate increasing amount of data.

1. **DESIGN**

**5.1 Introduction**

Number of modules: 3

Module1: Analysis based on age group of patients, gender, medication, etc.

Module2: Chi squared testing on different variables.

Module3: Linear Regression.

Module4: Linear Regression.

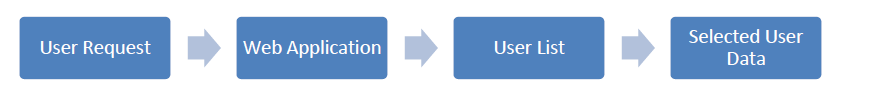
Algorithms: Clustering.(k-means Clustering)

**5.2 Architecture Design**

The system architecture follows the purpose and control flow of the application. This application is developed for extracting useful facts from unstructured data such as information regarding age group, gender and medical tests. In particularly, the users can get tips according to his or her personnel information.

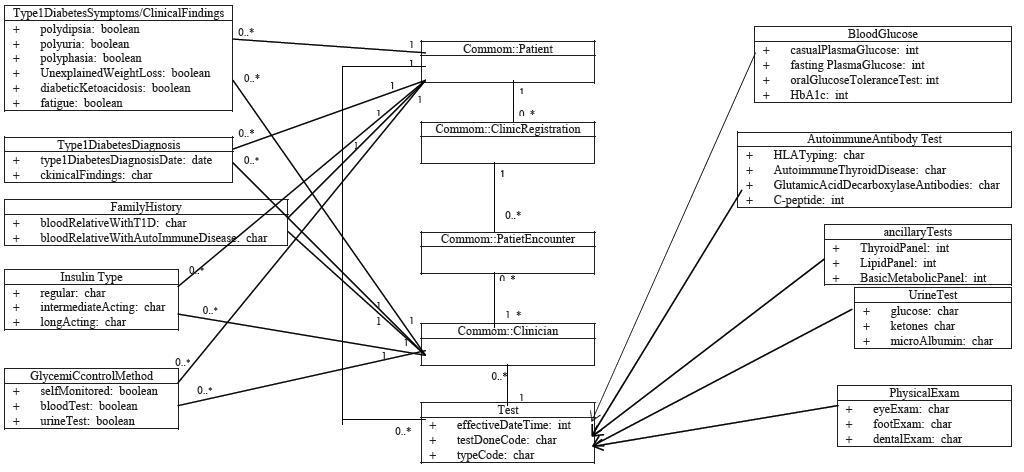


**User Control Flow**



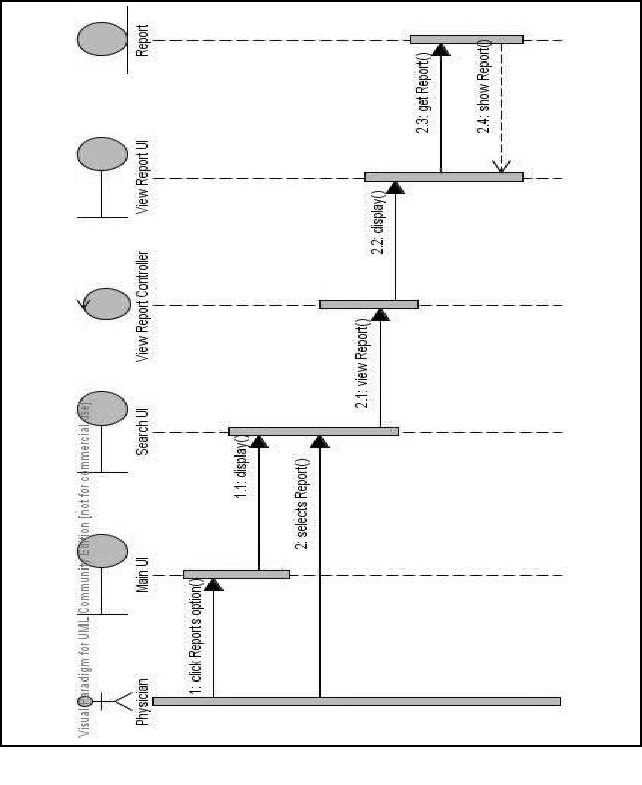
**Doctor Control Flow**

**5.3 Class Diagram:**

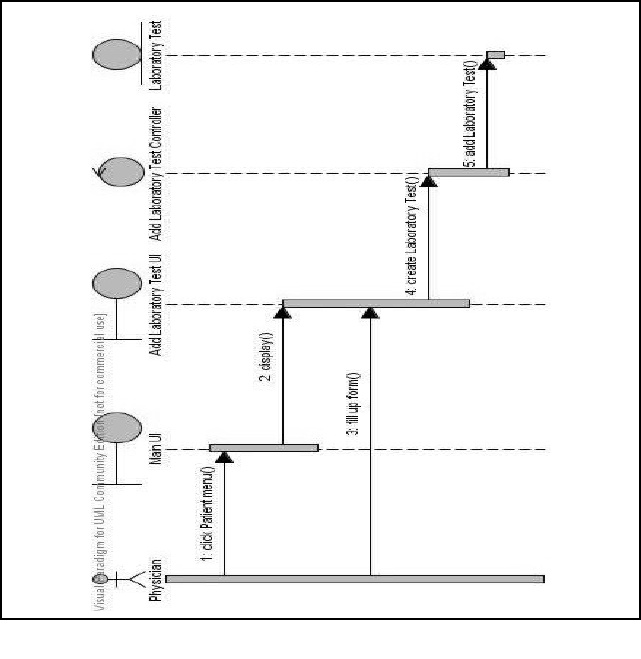


**5.4 Sequence Diagram:**

1. View Reports Sequence Diagram



1. Laboratory Test Sequence Diagram



1. **IMPLEMENTATION**

**6.1 Tools Introduction**

**R Studio** is a [free](https://en.wikipedia.org/wiki/Free_software) and [open source](https://en.wikipedia.org/wiki/Open-source_software) [integrated development environment](https://en.wikipedia.org/wiki/Integrated_development_environment) (IDE) for [R](https://en.wikipedia.org/wiki/R_%28programming_language%29), a [programming language](https://en.wikipedia.org/wiki/Programming_language) for [statistical computing](https://en.wikipedia.org/wiki/Statistical_computing) and graphics.

R Studio is available in two editions: R Studio Desktop, where the program is run locally as a regular [desktop application](https://en.wikipedia.org/wiki/Desktop_application); and R Studio Server, which allows accessing R Studio using a web browser while it is running on a remote [Linux](https://en.wikipedia.org/wiki/Linux) server. Prepackaged distributions of R Studio Desktop are available for [Microsoft Windows](https://en.wikipedia.org/wiki/Microsoft_Windows), [Mac OS X](https://en.wikipedia.org/wiki/Mac_OS_X), and Linux. R Studio is written in the [C++](https://en.wikipedia.org/wiki/C%2B%2B) programming language and uses the [Qt framework](https://en.wikipedia.org/wiki/Qt_%28software%29) for its [graphical user interface](https://en.wikipedia.org/wiki/Graphical_user_interface).

**6.2 Technology Introduction**

**R** is a [programming language](https://en.wikipedia.org/wiki/Programming_language) and software environment for [statistical computing](https://en.wikipedia.org/wiki/Statistical_computing) and graphics supported by the R Foundation for Statistical Computing. The R language is widely used among [statisticians](https://en.wikipedia.org/wiki/Statistician) and [data miners](https://en.wikipedia.org/wiki/Data_mining) for developing [statistical software](https://en.wikipedia.org/wiki/Statistical_software) and data analysis. Polls, [surveys of data miners](https://en.wikipedia.org/wiki/Rexer%27s_Annual_Data_Miner_Survey), and studies of scholarly literature databases show that R's popularity has increased substantially in recent years.

**6.3 Explanation of Algorithm and how it is been implemented**

**1. Analysis based on Insulin**

diabetic\_data <- read.csv("~/Desktop/dataset\_diabetes/diabetic\_data.csv")

i1<-subset(diabetic\_data,insulin=="Up")  
i2<-subset(diabetic\_data,insulin=="Down")  
i3<-subset(diabetic\_data,insulin=="Steady")  
i4<-subset(diabetic\_data,insulin=="No")  
x<-c(nrow(i1),nrow(i2),nrow(i3),nrow(i4))  
label<-c("Up", "Down", "Steady", "No")  
pie3D(x,labels=label,explode=0.1,main="Analysis based on Adminstration of Insulin")  
png(file="insulin.png")  
dev.off()  
png(file="insulin.png")  
pie3D(x,labels=label,explode=0.1,main="Analysis based on Adminstration of Insulin")  
dev.off()

**2. Analysis based on Medication**

i1<-subset(diabetic\_data,metformin="Steady")  
i2<-subset(diabetic\_data,glimepiride="Steady")  
i3<-subset(diabetic\_data,glipizide="Steady")  
i1<-subset(diabetic\_data,metformin=="Steady")  
i2<-subset(diabetic\_data,glimepiride=="Steady")  
i3<-subset(diabetic\_data,glipizide=="Steady")  
i4<-subset(diabetic\_data,glyburide=="Steady")  
i5<-subset(diabetic\_data,acarbose=="Steady")  
x<-c(nrow(i1),nrow(i2),nrow(i3),nrow(i4),nrow(i5))  
label<-c("Metformin", "Glimepiride","Glipizide", "Glyburide", "Acarbose")  
png(file="medicine.png")  
pie3D(x,labels=label,explode=0.1,main="Analysis based on Adminstration of Medication")  
dev.off()  
piepercent<- round(100\*x/sum(x),1)  
png(file="medicine\_final.png")  
pie(x,labels=piepercent,main="Analysis based on Administration of Medication",col=rainbow(length(x)))  
legend("topright",c("Metformin", "Glimepiride","Glipizide", "Glyburide", "Acarbose"),cex=0.8,fill=rainbow(length(x)))  
dev.off()

**3. Analysis based on Gender**

e <- max(diabetic\_data$encounter\_id)  
print(e)  
e <- max(diabetic\_data$age)  
f<- count(diabetic\_data, gender=="Female")  
f<- subset(diabetic\_data, gender=="Female")  
m<- subset(diabetic\_data, gender=="Male")  
fn<- nrow(f)  
mn<- nror(m)  
mn<- nrow(m)  
x<-c(fn,mn)  
labels<- c("Female","Male")  
png(file= "gender.jpg")  
pie(x,labels)  
dev.off()  
png(file= "gender1.jpeg")  
pie(x,labels)  
dev.off()

**4. Analysis based on Race**

View(diabetic\_data)

library(plotrix)  
r1<- subset(diabetic\_data, race=="Caucasian")  
r2<- subset(diabetic\_data, race=="AfricanAmerican")  
r3<- subset(diabetic\_data, race=="Asian")  
r4<- subset(diabetic\_data, race=="Hispanic")  
r5<- subset(diabetic\_data, race=="Other")  
x<- c(nrow(r1),nrow(r2),nrow(r3),nrow(r4),nrow(r5))  
label<- c("Caucasian","African American","Asian","Hispanic","Other")  
png(file="race\_final.png")  
pie(x,labels=piepercent,main="Analysis based on Race",col=rainbow(length(x)))  
legend("topright",c("Caucasian","AfricanAmerican","Asian","Hispanic","Others"),cex=0.8,fill=rainbow(length(x)))  
dev.off()

**5. Chi Square test**

library("MASS")

diabetes.data = table(diabetic\_data$gender, diabetic\_data$change)

print(diabetes.data)

print(chisq.test(diabetes.data))

**6. Logistic Regression based on gender and HbA1c results**

plot(diabetes4$A1Cresult, jitter(diabetes4$gender, 0.15),pch=19,xlab="Primary Diagnosis", ylab="Sex(1-Female, 2-Male)")

diabetes5$A1Cresult<- as.numeric(diabetes5$A1Cresult)

xv<- seq(min(diabetes5$A1Cresult),max(diabetes5$A1Cresult),0.01)

model<- glm(diabetes5$gender~diabetes5$A1Cresult,binomial)

yv<-predict(model,list(xv),type="response")

lines(xv,yv,col="red")

**7. Logistic Regression based on gender and Primary Diagnosis**

plot(diabetes4$diag\_1 , jitter(diabetes4$gender, 0.15),pch=19,xlab="Primary Diagnosis", ylab="Sex(1-Female, 2-Male)")

**8. Linear Regression:**

relation<- lm(diabetes5$gender~diabetes5$A1Cresult)

print(relation)

png(file="linear model.png")

plot(diabetes5$A1Cresult,diabetes5$gender,col="red",main="Gender and HbA1c result regression",abline(lm(diabetes5$gender~diabetes5$A1Cresult)),cex=1.3,pch=16,xlab="HbA1c result",ylab="Gender")

dev.off()

**9. k-means clustering**

result<- kmeans(diabetes6, 4)

result

plot(diabetes6, col = result$cluster)

**10. K-means clustering result:**

K-means clustering with 4 clusters of sizes 10697, 25353, 5771, 28166

Cluster means:

race gender age A1Cresult

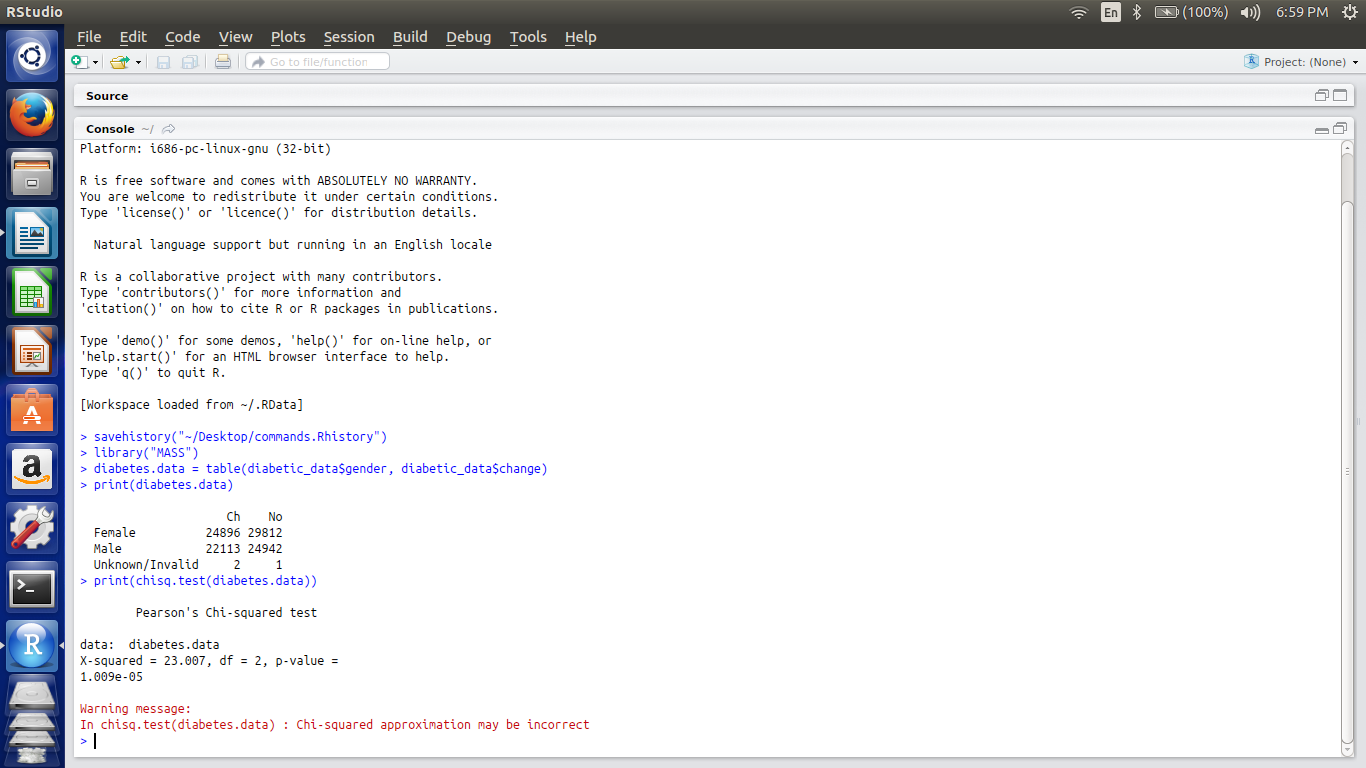
1 2.465551 0.4674208 3.342526 3.685052

2 2.335660 0.5115765 5.561866 3.900327

3 2.327326 0.4805060 6.363022 1.403396

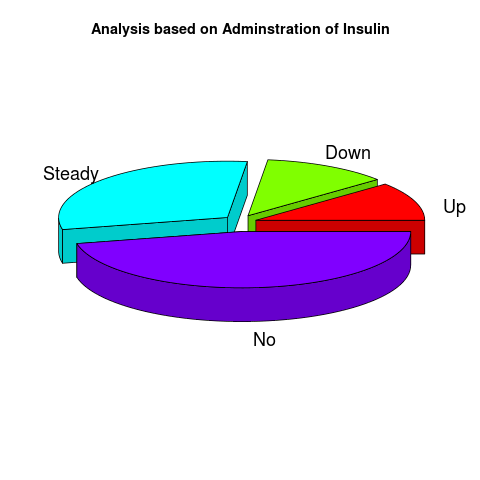
4 2.193709 0.4262941 7.492828 3.943336

**11.Chi-Square**

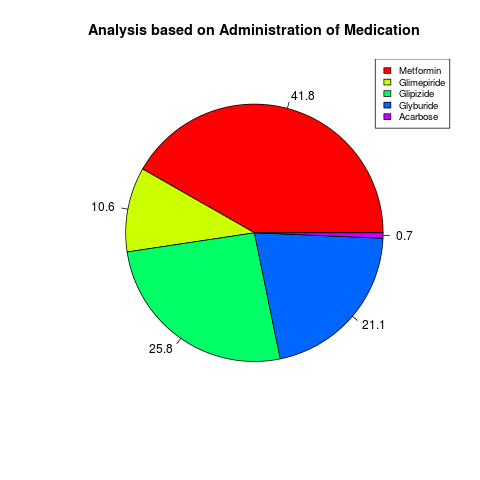


1. **TESTING**

**Figure 1.** A 3D pie chart representing analysis of administration of insulin

****

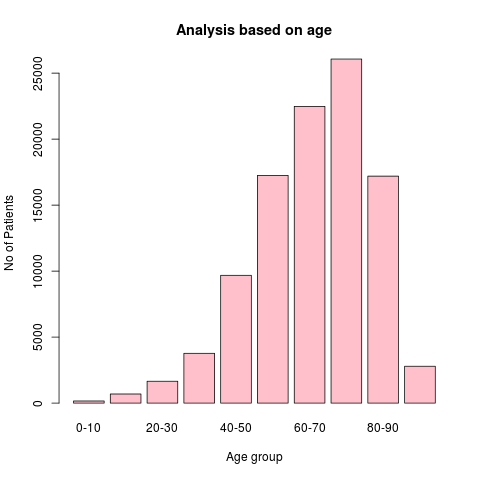
**Figure 2.** A pie chart representing the analysis based administration of medication

****

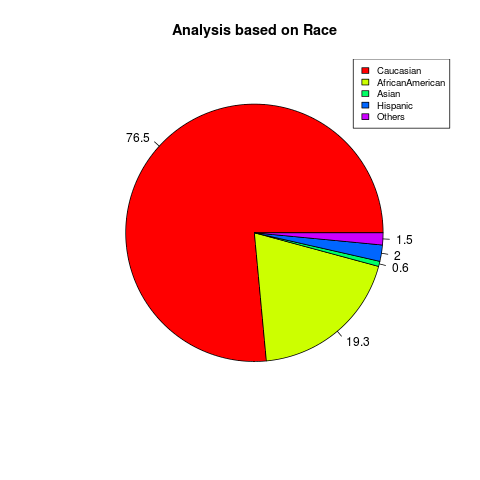
**Figure 3.** A 3D pie chart showing distribution based on gender

****

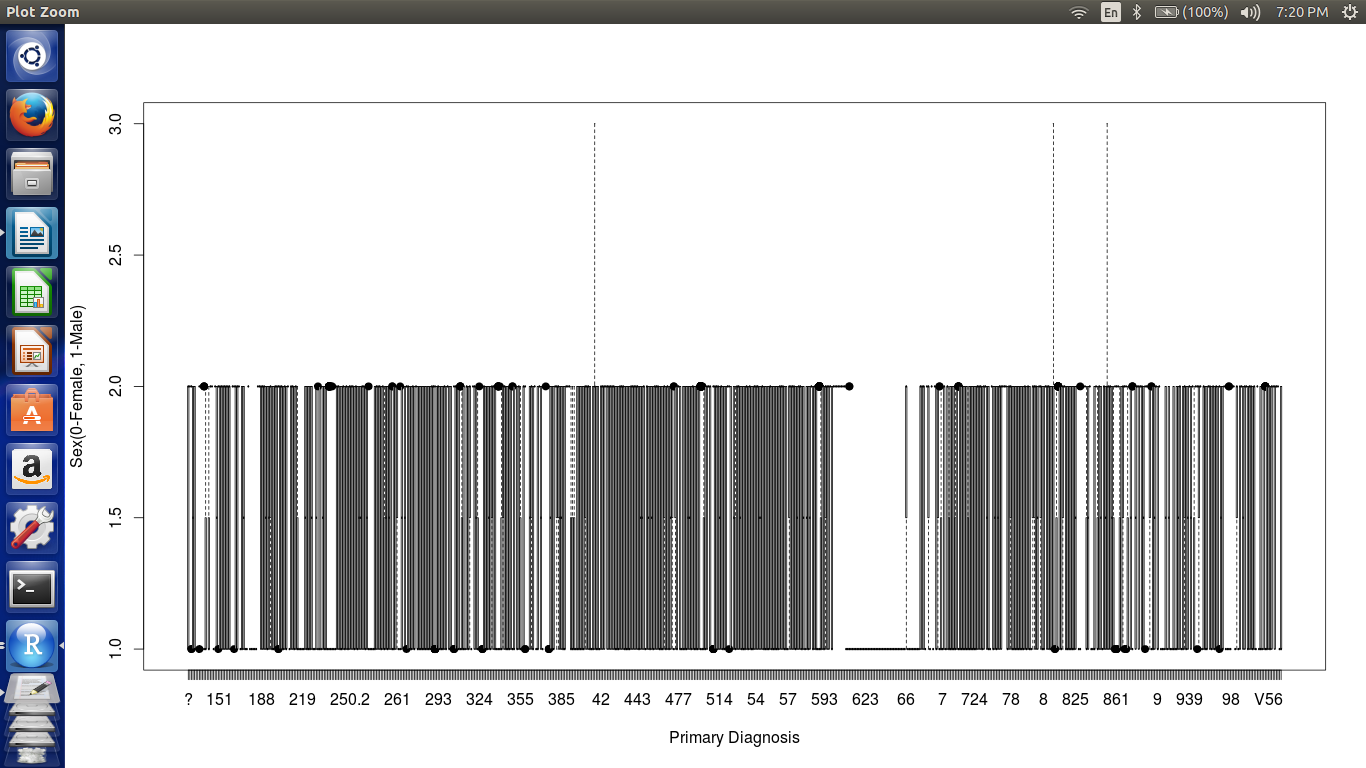
**Figure 4.** A bar graph showing distribution of number of patients based on age group in intervals of 10 years

****

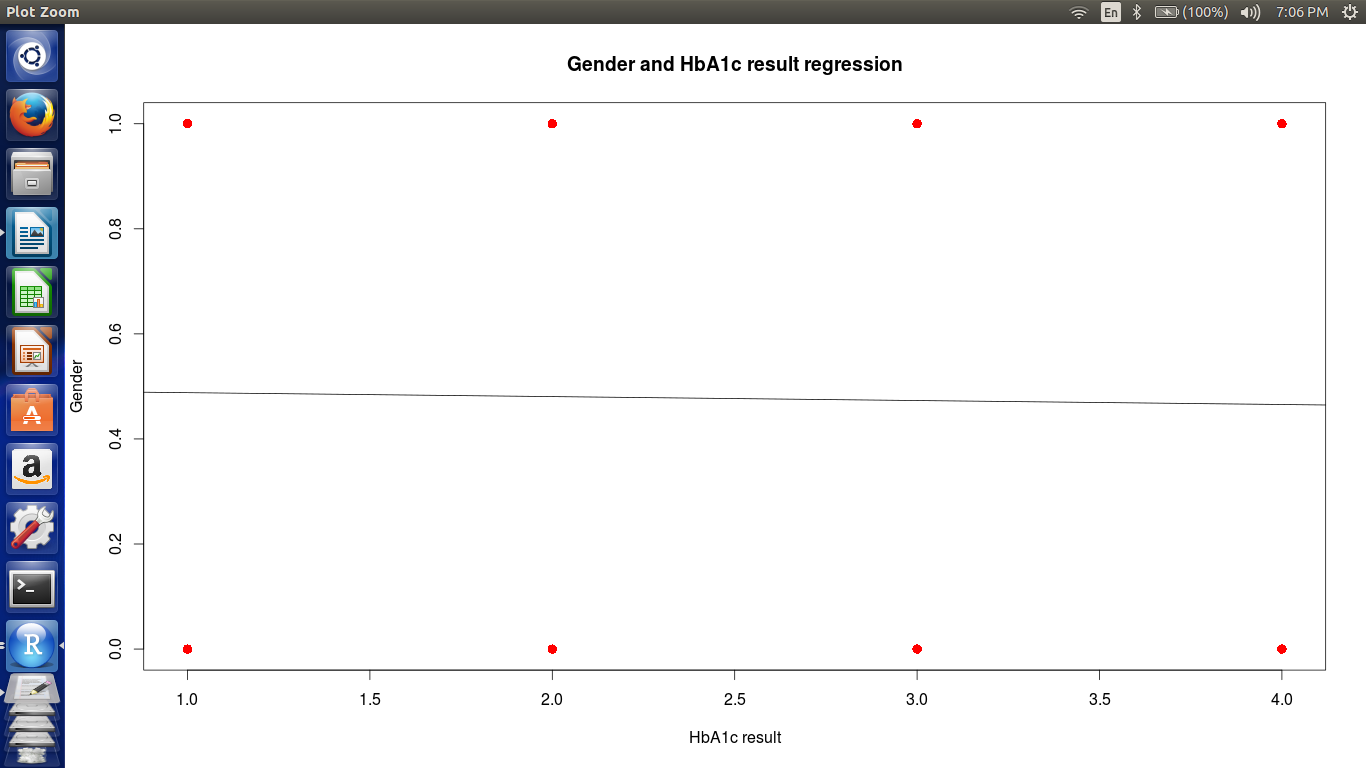
**Figure 5.** Analysis based on race

****

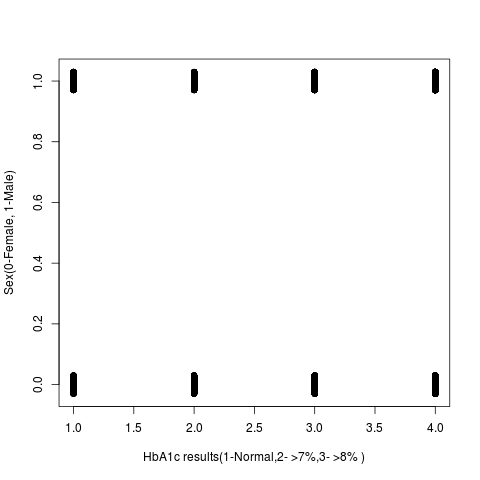
**Figure 6.**  A logistic regression showing relation between gender and primary diagnosis test.

****

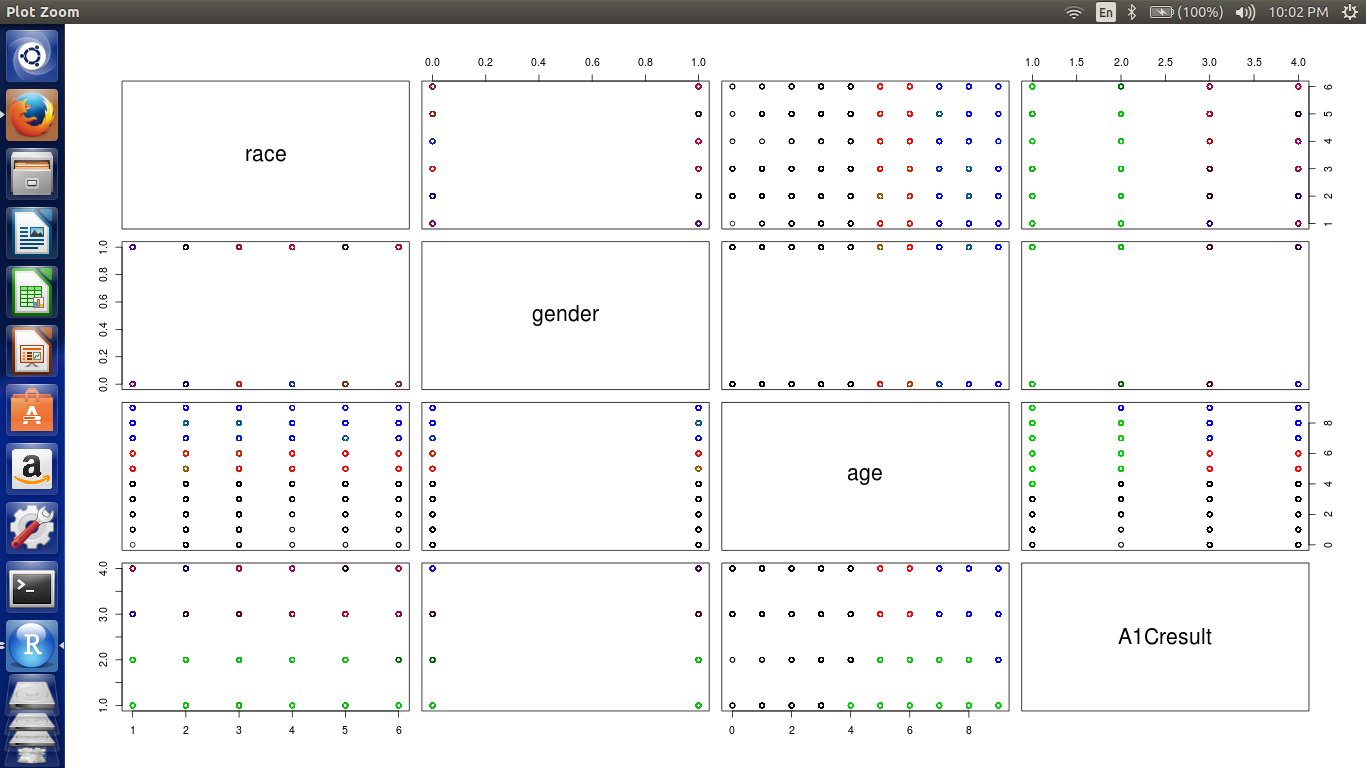
**Figure 7.** A linear regression showing relation between gender and HbA1c test.

****

**Figure 8.** A logistic regression showing relation between gender and HbA1c test.

****

**Figure 9.** A k- means clustering on four variables(race,gender,age,HbA1c results)

****

1. **CONCLUSION**

To prevent diabetes related morbidity and mortality, there is an immense need of dedicated self-care behaviors in multiple domains, including food choices, physical activity, proper medications intake and blood glucose monitoring from the patients. Though multiple demographic, socio-economic and social support factors can be considered as positive contributors in facilitating self-care activities in diabetic patients, role of clinicians in promoting self-care is vital and has to be emphasized. Realizing the multi-faceted nature of the problem, a systematic, multi-pronged and an integrated approach is required for promoting self-care practices among diabetic patients to avert any long-term complications.

1. **REFERENCES**

1. <http://clinical.diabetesjournals.org/content/29/3/102.full>

2.[http://randyzwitch.com/hive-five-hard-won-lessons/](http://randyzwitch.com/hive-five-hard-won-lessons/" \t "_blank)  
3.http://randyzwitch.com/big-data-hadoop-amazon-ec2-cloudera-part-1

4.http://jdmdonline.biomedcentral.com/articles/10.1186/2251-6581-12-14

5.[http://www.gardenersown.co.uk/education/lectures/R/index.htm](http://www.gardenersown.co.uk/education/lectures/R/index.htm" \t "_blank)

6.<https://archive.ics.uci.edu/ml/datasets/Diabetes+130-US+hospitals+for+years+1999-2008>