**Executive Summary for Project on Diabetes**

1. What is the problem you want to solve:

Diabetes is a chronic disease that occurs either when the pancreas does not produce enough insulin or when the body cannot effectively use the insulin it produces leading to serious damage to many of the body's systems, especially the nerves and blood vessel.  The number of people with diabetes has risen from 108 million in 1980 to 422 million in 20141.

The global prevalence of diabetes among adults over 18 years of age has risen from 4.7% in 1980 to 8.5% in 20141. Diabetes is a major cause of blindness, kidney failure, heart attacks, stroke and lower limb amputation1. While individual treatments is available for diabetes, insights from large chuck of data helps in prevention & identify people with high risk factors for Diabetes. At the same time diabetes can be prevented by lifestyle changes which will help in improving insulin sensitivity there by preventing Diabetes. Prediabetes means that your blood sugar level is higher than normal but not yet high enough to be classified as type 2 diabetes. Without intervention, prediabetes is likely to become type 2 diabetes in 10 years or less.2 Mayoclinic also states that certain lifestyle changes including eating healthy foods, getting more physical activity &loosing excess pounds can treat or even change they can reverse their condition. Hence to identify patients who are at risk of diabetes – this paper would like to explore and identify different markers of diabetes and identify the following

* Create a model leverage early detection of Diabetes /increase predictability of diabete
* Identify similarities in patients who have good sugar control as compared to other diabetic patients
* Identify other markers of this disease.

Benefits of Research will help following clients

|  |  |  |
| --- | --- | --- |
| S.No | Type of Client | How they could benefit from my Research |
| 1 | Insurance Companies | Increase Insurance cost for candidates who have higher predictability of Diabetes rationalizing revenue and prediction model for its earnings |
| 2 | Health Companies/Hospitals | Better Cognition and advance detection of disease leading to better monitoring of patients |

1. Deep Dive of my Data Set:

* What important fields and information does the data set have?
* What are its limitations i.e. what are some questions that you cannot answer with this data set?
* What kind of cleaning and wrangling did you need to do?

**Important Fields & information this data has:**

Source of Data: <https://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes>

In particular, all patients here are females’ at least 21 years old of Pima Indian heritage. The breakdown of various data entities are as under:

1. Number of times pregnant

2. Plasma glucose concentration a 2 hours in an oral glucose tolerance test

3. Diastolic blood pressure (mm Hg)

4. Triceps skin fold thickness (mm)

5. 2-Hour serum insulin (mu U/ml)

6. Body mass index (weight in kg/(height in m)^2)

7. Diabetes pedigree function

8. Age (years)

9. Class variable (0 or 1) where 0 means No Diabetes and 1 are Diabetic Group

**Limitations of the data**

The data in the dataset has only females and hence the marker for diabetes is useful only for females. The data in this set only has females of one origin. Hence multi facet analysis on ethnicity cannot be done. Certain data points have zero in BMI, Diastolic Pressure which is not biologically possible. These have been considered as outliers and have been removed from the analysis.

Another important factor which is not considered in the data set is amount of physical exercise performed. This is an important parameter which could help research substantiate on importance of exercise regimen on general health.

**Data Cleaning & Wrangling**

The dataset did not have any headers and hence as a first step headers were created using following statement in R

names(pima.indians.diabetes)<-c('TimesPregnant','GlucoseLV','DiastolicBP','TricepsThickness','SerumInsulin','BMI','Heridarymarkup','Age','Classification')

The dataset had 0 biological values in certain fields like Diastolic Pressure, BMI, Triceps Thickness, and Serum Insulin. Hence we needed to remove Outliers by considering 0 values as NA.0 values were replaced by NA.

pima.indians.diabetes$DiastolicBP[pima.indians.diabetes$DiastolicBP==0]<-NA

pima.indians.diabetes$GlucoseLV[pima.indians.diabetes$GlucoseLV==0]<-NA

pima.indians.diabetes$TricepsThickness[pima.indians.diabetes$TricepsThickness==0]<-NA

pima.indians.diabetes$SerumInsulin[pima.indians.diabetes$SerumInsulin==0]<-NA

pima.indians.diabetes$BMI[pima.indians.diabetes$BMI==0]<-NA

pima.indian.diabetes2<-na.omit(pima.indians.diabetes)

4In brief, outline your approach to solving this problem (knowing that this might change later)

In order to compare Diabetic Group and Non Diabetic Group- another column based to differentiate Normal group and Diabetic group. Similarly to enhance the data we created factors called Obese vs Normal BMI & High Frequency Pregnancy vs Normal Pregnancy.

pima.indian.diabetes4<-pima.indian.diabetes2%>%mutate(Group=ifelse(pima.indian.diabetes2$Classification==1,"Diabetes","Normal"))

pima.indian.diabetes7<-pima.indian.diabetes2%>% mutate(BMIfactor=ifelse(pima.indian.diabetes2$BMI>27,"Overweight","Normal"))

pima.indian.diabetes8<-pima.indian.diabetes7%>% mutate(PregGroup=ifelse(pima.indian.diabetes7$TimesPregnant>3,"High","Normal"))

Also as a part of research have create a data set to compare subjects with Normal

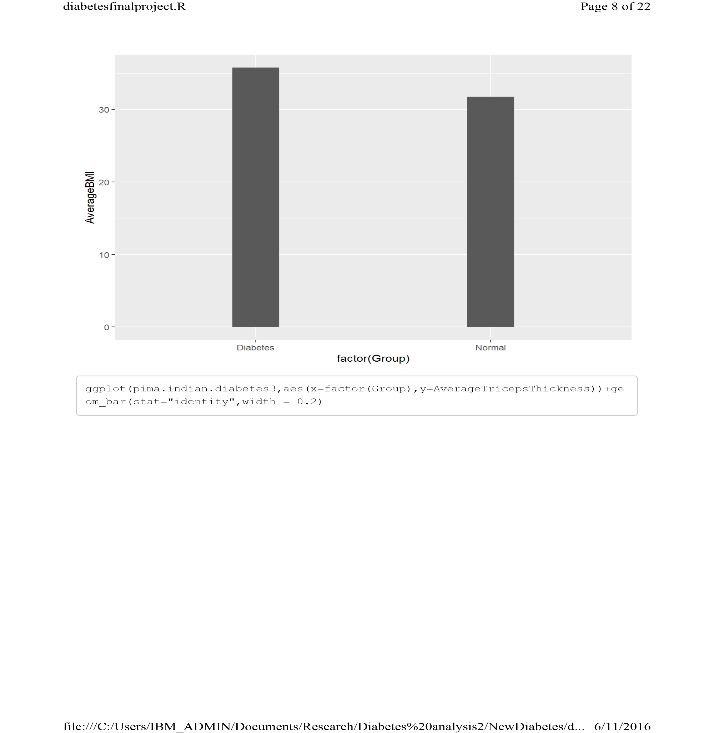
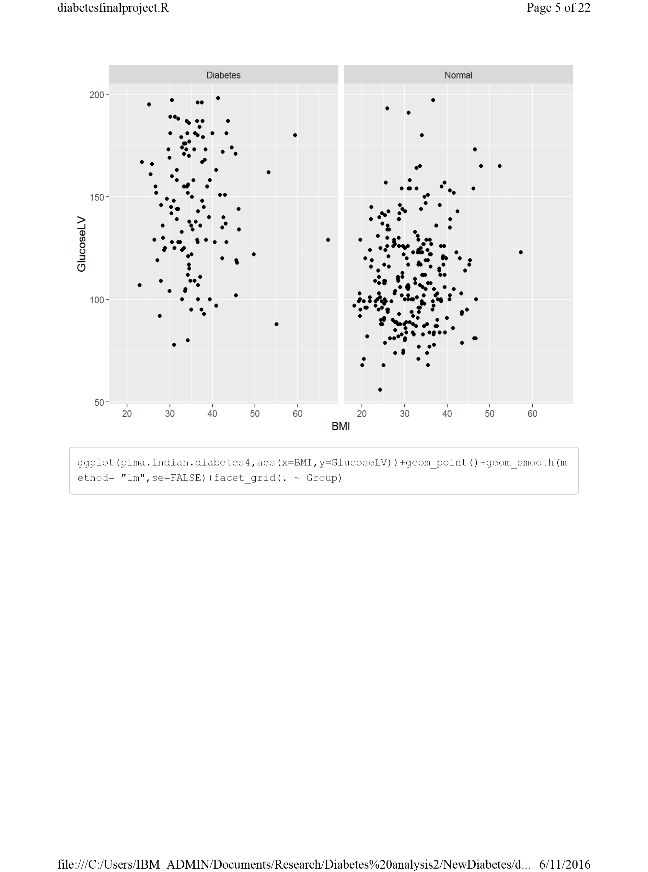
**Prediabetes<-pima.indian.diabetes8%>% filter(GlucoseLV>140& GlucoseLV<200 & Classification==0)**

**controlleddiabetes<-pima.indian.diabetes8%>% filter(GlucoseLV<140 & Classification==1)**

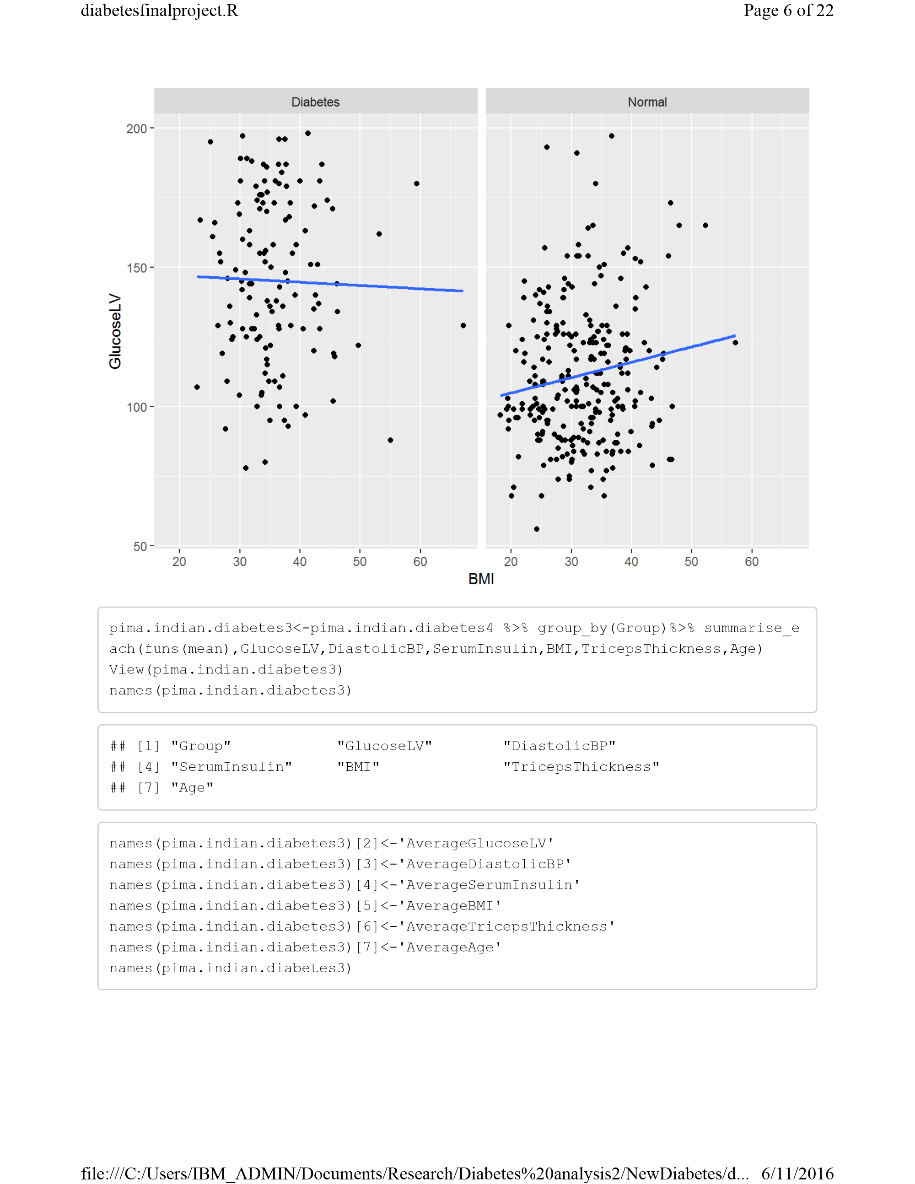
**Normal<-pima.indian.diabetes8%>%filter(GlucoseLV<140 & Classification==0)**

**Initial Findings**

The objective of initial findings was to compare the key parameters and their difference between Normal Group & Diabetic Group.Preliminary findings suggest that Diabetic subjects have higher BMI as compared to Normal Group

** **

For normal subjects- BMI has a positive correlation with Plasma (2 hour Post Meal Glucose). However for diabetic patients this is not true as expected because all of them will be having medications to control their glucose level

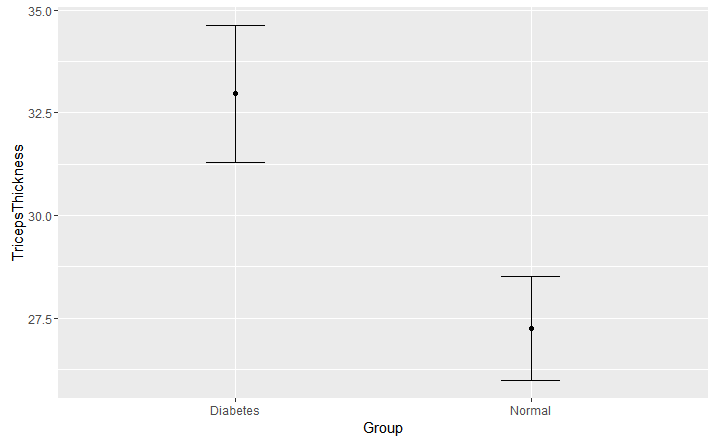
****

mean /average were plotted with different predictors in both diabetic & Normal Group

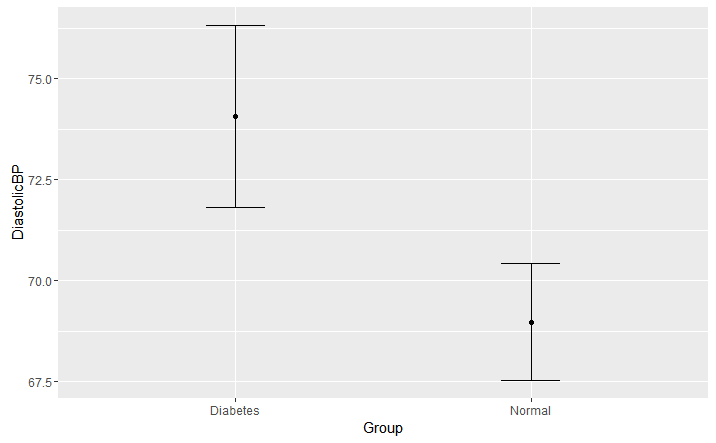
1. Mean & error with confidence 95% among three existing factors for Diabetes namely- BMI,Tricep Thickness & Diastolic Blood Pressure

All the below graph shows that Diabetic people have higher BMI,greater Triceps thickness & Higher Diastolic Pressure

**Triceps thickeness between control Groups**

****

**Diastolic Pressure between control groups**

****

**BMI between control groups**

****

**Initial Modelling Findings**

Creating a Mathematical model using Logistic Regression. We would need a logistic regression because independent variable is either continuous or factor/ Discrete while Dependent variable is binary (0- No Diabetes & 1 Diabetes)

**Mathematical Approach:**

…………………EQUATION 1

Leverage the above formula y (Diabetes) for different subsets by creating different subsets & factors like High BMI, Low BMI, Triceps Thickness. Using Natural Log

……………………..Equation 2

set.seed(88)

split<-sample.split(pima.indian.diabetes8$Classification,SplitRatio =0.75)

split

pima.indian.diabetestrain<-subset(pima.indian.diabetes8,split==TRUE)

pima.indian.diabetestest<-subset(pima.indian.diabetes8,split==FALSE)

nrow(pima.indian.diabetestrain)

nrow(pima.indian.diabetestest)

View(pima.indian.diabetestrain)

summary(pima.indian.diabetestrain)

Trainlog<-glm(Classification~TricepsThickness+factor(BMIfactor)+DiastolicBP,data=pima.indian.diabetestrain,family=binomial)

summary(Trainlog)

predictTrain<-predict(Trainlog,type ="response",data=pima.indian.diabetestrain)

**Model 1 with independent variable (BMI) as Factor while other factors are continous**

**Summary Result**

Call:

glm(formula = Classification ~ TricepsThickness + BMI + DiastolicBP,

family = binomial, data = pima.indian.diabetestrain)

Deviance Residuals:

Min 1Q Median 3Q Max

-1.6724 -0.8828 -0.6676 1.2292 2.0112

Coefficients:

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

(Intercept) -4.38257 0.91402 -4.795 1.63e-06 \*\*\*

TricepsThickness 0.02709 0.01656 1.636 0.1019

BMI 0.05312 0.02449 2.169 0.0301 \*

DiastolicBP 0.01506 0.01107 1.361 0.1737

---

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

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 374.27 on 293 degrees of freedom

Residual deviance: 347.13 on 290 degrees of freedom

AIC: 355.13

Number of Fisher Scoring iterations: 4

**Probabilistic Modelling**

tapply(predictTrain,pima.indian.diabetestrain$Classification,mean)

0 1

0.304115 0.391770

> tapply(predictTrain,pima.indian.diabetestrain$BMIfactor,mean)

Normal Overweight

0.1206897 0.3855932

> table(pima.indian.diabetestrain$Classification,predictTrain>0.35)

FALSE TRUE

0 114 82

1 35 63

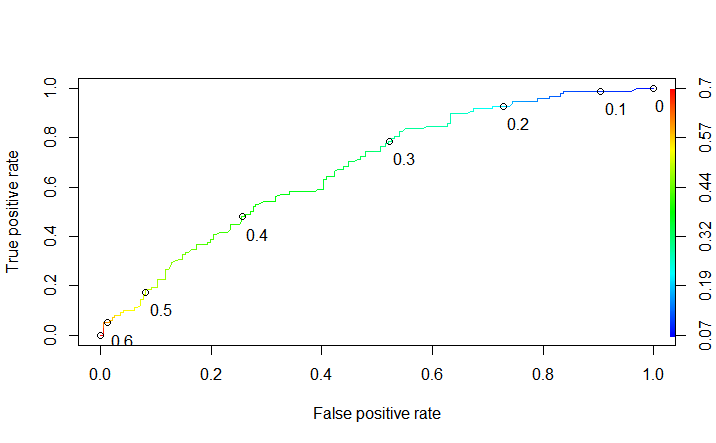
**Hence based on my initial modelling the Bayesian Matrix/ Confusion matrix would show as following:**

**Mean Probability of Normal BMI getting diabetes=p(BMI=Normal)= 0.12**

**Mean Probability of Overweight=p(BMI=Overweight)=0.391**

|  |  |  |  |
| --- | --- | --- | --- |
| **Actual Values** | **Prediction= No Diabetes** | | **Prediction= Diabetes** |
| **No Diabetes** | **114** | **82** | |
| **Diabetes** | **35** | **63** | |

**Modelling of Sensitivity & Specificity against different Probabilistic models:**

****

**Model 2: Taking all x variables as continuous and without any interactions.**

tapply(predictTrain2,pima.indian.diabetestrain$Classification,mean)

0 1

0.304273 0.391454

> table(pima.indian.diabetestrain$Classification,predictTrain2>0.25)

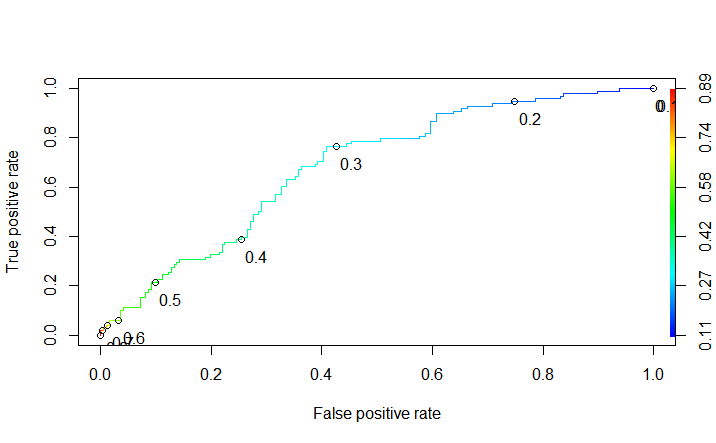
FALSE TRUE

0 79 117

1 17 81

|  |  |  |
| --- | --- | --- |
| **Actual Values** | **Prediction= No Diabetes** | **Prediction= Diabetes** |
| **No Diabetes** | **79** | **117** |
| **Diabetes** | **17** | **81** |

**Rates for given probability level**

****