Sugar Rush : Prediction and Analysis of Diabetes in PIMA Women using Logistic Regression

ROLL NO - 21BDA59

NAME - Purva Singh

Team Name - The Outliers

Contents

Link of presentation - Sugar Rush _ Diabetes Prediction.pdf

Link of screen recording of dashboard - https://drive.google.com/file/d/1mGnYDJZ0fsv7SDYePe6yHCWxdUKFWXDU/view?usp=sharing

Introduction

- Diabetes is a chronic (long-lasting) health condition that affects how your body turns food into energy. Most of the food you eat is broken down into sugar (also called glucose) and released into your bloodstream.
- When your blood sugar goes up, it signals your pancreas to release insulin. Insulin acts like a key to let the blood sugar into your body's cells for use as energy.
- If you have diabetes, your body either doesn't make enough insulin or it cannot use the insulin it makes as well as it should. When there isn't enough insulin or cells stop responding to insulin, too much blood sugar stays in your bloodstream. Over time, that can cause serious health problems, such as heart, kidney disease.

Why did we choose this topic and why did we only focus only on women?

Diabetes affects women and men in almost equal numbers. However, diabetes affects women differently than men. Compared with men with diabetes, women with diabetes have.

- A higher risk for heart disease. Heart disease is the most common complication of diabetes
- A higher risk for blindness
- A higher risk for depression. Depression, which affects twice as many women as men, also raises the risk for diabetes in women

Problem Statement

The goal of this project is to build a logistic regression model that would predict the likelihood of diabetes and perform analysis on the risk factors, particularly in women, the PIMA Indians' Diabetes dataset was chosen.

Data Methodology

1) Data Collection -

- The diabetes data contains information about PIMA Indian females, due to the high incidence rate of Diabetes in PIMA females.
- The dataset was originally published by the National Institute of Diabetes and Digestive and Kidney Diseases, consisting of diagnostic measurements pertaining to females of age greater than 20.
- It contains information of 768 females, of which 268 females were diagnosed with Diabetes. Information available includes 8 variables, such as, Age, Number of Pregnancies, Glucose, Insulin, etc.

Factors taken into consideration

Pregnancies	Number of times a person is pregnant
Glucose	Plasma glucose concentration over 2 hours in an oral glucose tolerance test
Blood Pressure	Diastolic blood pressure (mm Hg)
Skin Thickness	Triceps skin fold thickness (mm)
Insulin	Helps in breakdown so that glucose can be circulated throughout cells
BMI	Body mass index (weight in kg/(height in m)2)
Diabetes Pedigree Function	Diabetes pedigree function (a function which scores likelihood of diabetes based on family history)
Age	Age (years)

2) **DATA PREPROCESSING**

- We observed that in the data there are no null values
- But there were some independent variables which has 0 value.
- Blood pressure , BMI, Insulin, Skin Thickness, Glucose had 0 values but these factors cannot have 0 values
- So we used K nearest neighbor imputation to replace the 0 values

USER INTERFACE DEVELOPMENT

1) Data Exploration -

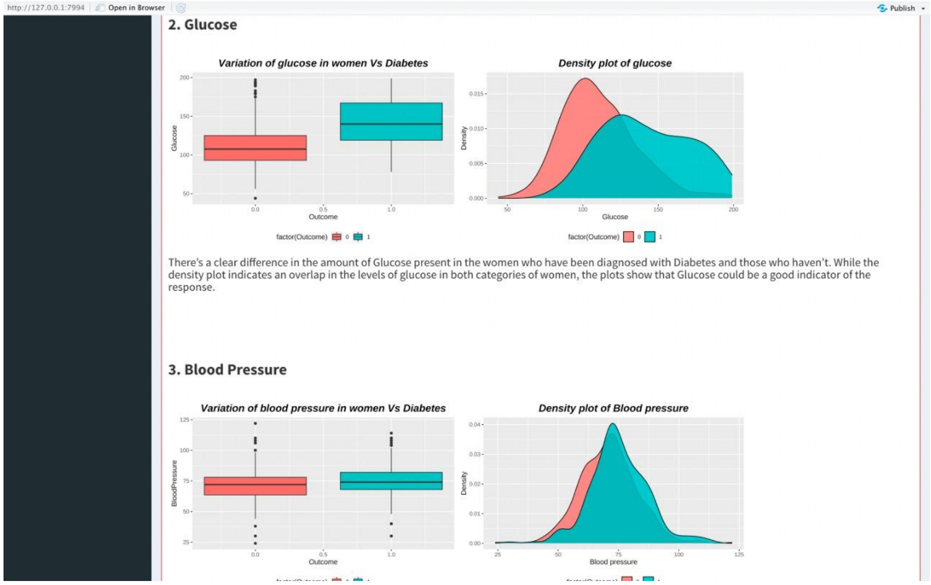
In this tab we basically analyzed how much good or bad that variable is for the prediction,i.e,Variable VS Outcome (outcome is response variable and others are predictor variable) -

After our analysis, we found that

- though there's some difference in the levels of boxplot of all variables except Glucose but due to overlapping in density plot , we can say that those variables isn't a good factor when while predicting

Glucose is good factor while predicting as -

- there's less overlapping and significant difference between levels of boxplot, so we can say that



2. Glucose

Variation of glucose in women Vs Diabetes

Density plot of glucose

There's a clear difference in the amount of Glucose present in the women who have been diagnosed with Diabetes and those who haven't. While the density plot indicates an overlap in the levels of glucose in both categories of women, the plots show that Glucose could be a good indicator of the response.

3. Blood Pressure

Variation of blood pressure in women Vs Diabetes

Density plot of Blood pressure

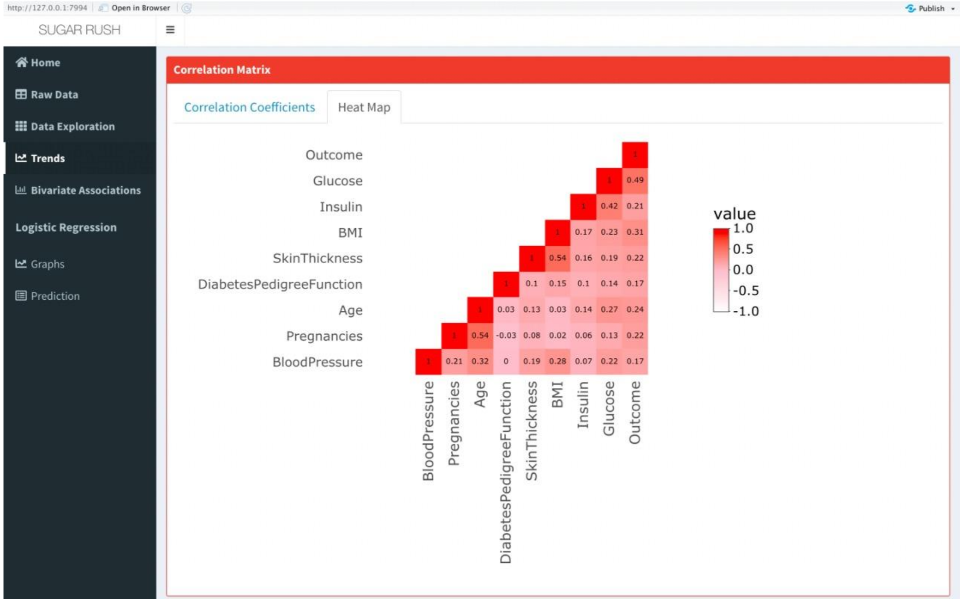
2) <u>Heat Map -</u>

A heat map (or heatmap) is a graphical representation of data where values are depicted by color. Heat maps make it easy to visualize complex data and understand it at a glance.

Here from our heatmap we can see that **highest correlation** is between

- Skin Thickness and BMI
- Age and Pregnancy

And when it comes to comparing with **Outcome, Highest correlation** is between **Glucose and Outcome.**



Correlation Matrix

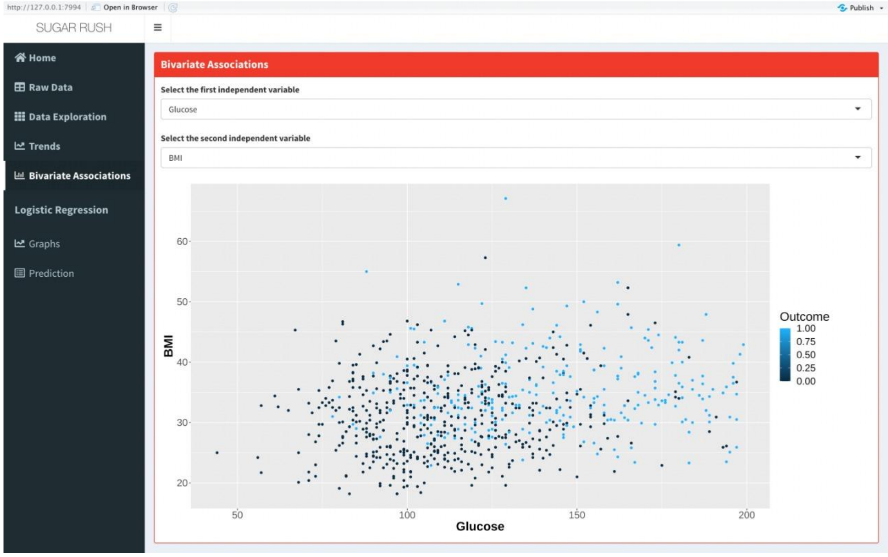
	BloodPressure	Pregnancies	Age	DiabetesPedigreeFunction	SkinThickness	BMI	Insulin	Glucose	Outcome
Outcome									1
Glucose								1	0.49
Insulin							1	0.42	0.21
BMI						1	0.17	0.23	0.31
SkinThickness					1	0.54	0.16	0.19	0.22
DiabetesPedigreeFunction				1	0.1	0.15	0.1	0.14	0.17
Age			1	0.03	0.13	0.03	0.14	0.27	0.24
Pregnancies		1	0.54	-0.03	0.08	0.02	0.06	0.13	0.22
BloodPressure	1	0.21	0.32	0	0.19	0.28	0.07	0.22	0.17

3) - Bivariate Associations Tab

Bivariate analysis is the simultaneous analysis of two variables (attributes). It explores the concept of relationship between two variables, whether there exists an association and the strength of this association, or whether there are differences between two variables and the significance of these differences.

For example, this graph is showing that at a specific value of Glucose and BMI, the person is having diabetes or not-

- if the glucose is between 50 and 100 and BMI is between 20-40 , the graph is representing that person does not have diabetes wheres if the glucose is greater than 100 and BMI is between 30-50, there’s a high chance that person can get diabetes



Select the first independent variable

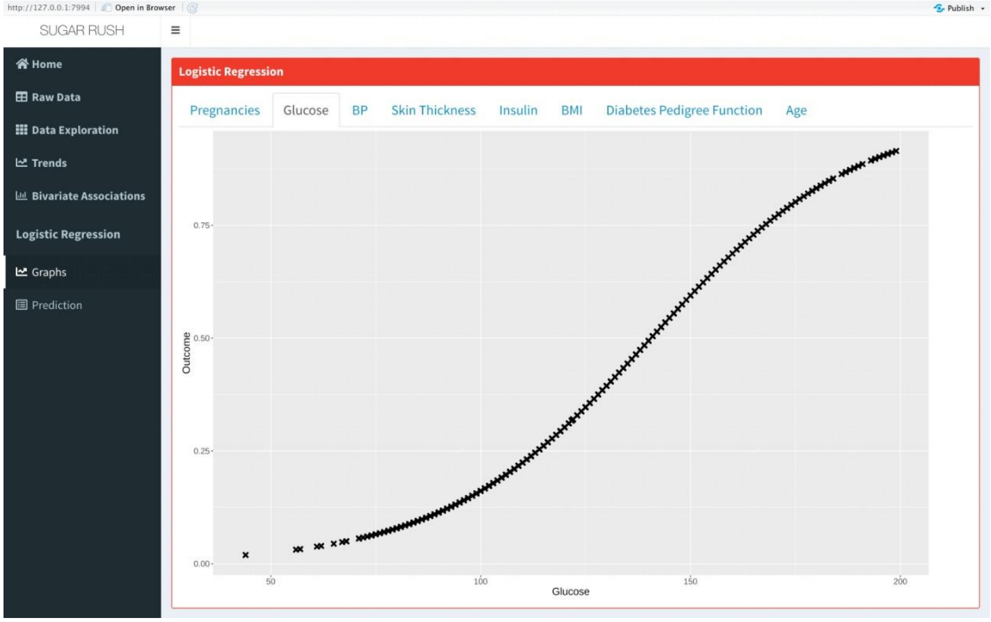
Glucose

Select the second independent variable

BMI

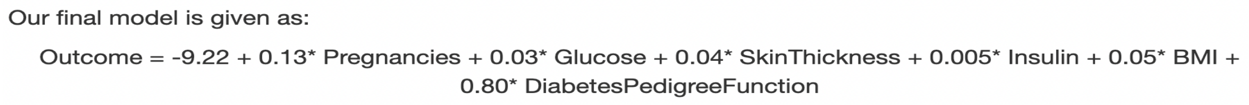
4) - Logistics Regression/Prediction

For our dataset, we have built a Logistic Regression Model as the response variable is a binary classifier which can hold only two values 0 and 1. Logistic regression has the ability to provide both the probabilities and classify them using continuous and discrete data. In our case, if the probability is greater than 0.5, the patient is classified as diabetic and non-diabetic otherwise. Unlike in linear regression, in logistic regression, We are fitting an S shaped logistic function.



For the model we consider only those variables whose effect on prediction is significant.

Based on the p-values, we have eliminated Age and Blood pressure from the model.



Our final model is given as:

Outcome = -9.22 + 0.13* Pregnancies + 0.03* Glucose + 0.04* SkinThickness + 0.005* Insulin + 0.05* BMI + 0.80* DiabetesPedigreeFunction

Conclusion-

- The PIMA Women’s Database was analyzed and explored in detail. The patterns identified using Data exploration methods were validated using the modeling techniques like logistics regression
- By analysis and modeling, we observed that there’s no significant correlation between factors. Glucose has the maximum correlation wrt Outcome.Highest correlation is between Age & Pregnancy and Skin Thickness & BMI.
- The patterns identified using Data
- exploration methods were validated using the logistic regression model. The model is also used to predict whether the patient is diabetic or not depending on the user input values.
- We have created an interactive dashboard using R Shiny, HTML and CSS to display the entire analysis and prediction. The model has an accuracy of 76%.

Future Work -

In the future, we will work on improving the accuracy of the model for better performance by expanding the dataset and adding more significant variables. The user interface can be made more interactive and visually pleasing for the users.

References -

1) www.kaggle.com
2) https://www.who.int/
3) https://en.wikipedia.org/wiki/Pima_people