Sugar Rush :

Prediction and Analysis of Diabetes in Pima Women using Logistic Regression

OUTLIER : ARYA CHANDRAN

REG NO : 21BDA34

CLASS : 1ST MSc BIG DATA ANALYTICS

TABLE OF CONTENTS

ABSTRACT

Due to the rapid increase in the number of diabetic people, it is essential to rightly identify the factors that contribute to the occurrence of diabetes. In this project, we have focused solely on diabetes in Pima women. The objective is to create an interactive dashboard using R Shiny, HTML and CSS to display the analysis on the data and build a logistic regression model to predict if the patient is diabetic or not. Overall, the model gave an accuracy of 76%.

INTRODUCTION

Diabetes is a chronic health condition that affects how your body turns food into energy. Over time, that can cause serious health problems, such as heart disease, kidney disease etc.

Due to increasing incidence rate of diabetes and prediabetes, it is a pressing issue in the health care industry.

Diabetes affects women and men in almost equal numbers. However, diabetes affects women more than men as women have higher risk for heart disease, blindness, depression etc.

The project focuses on Pima females which is a group of Native Americans living in an area consisting of what is now central and southern Arizona, as well as northwestern Mexico in the states of Sonora and Chihuahua.

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.

METHODOLOGY

DATA COLLECTION

The diabetes dataset was collected from Kaggle. It includes diagnostic measurements pertaining to PIMA women of age greater than 20 located near Phoenix, Arizona. It has been under continuous study since 1965 due to the high incidence rate of Diabetes in PIMA females. It was originally published by the National Institute of Diabetes and Digestive and Kidney Diseases.

It contains information of 768 females, of which 268 females were diagnosed with Diabetes. Information available includes 8 predictor variables and 1 response variable.

Variable Name	Variable Description
Pregnancies	Number of times pregnant
Glucose	Plasma glucose concentration at 2 hours in an oral glucose tolerance test
BloodPressure	Diastolic blood pressure
SkinThickness	Triceps skin fold thickness
Insulin	2-hour serum insulin (μU/ml)
BMI	Body Mass Index
DiabetesPedigreeFunction	Synthesis of the history of Diabetes Mellitus in relatives, generic
relationship of those relat	subject
Age	Age of the individual
Outcome	Occurrence of Diabetes

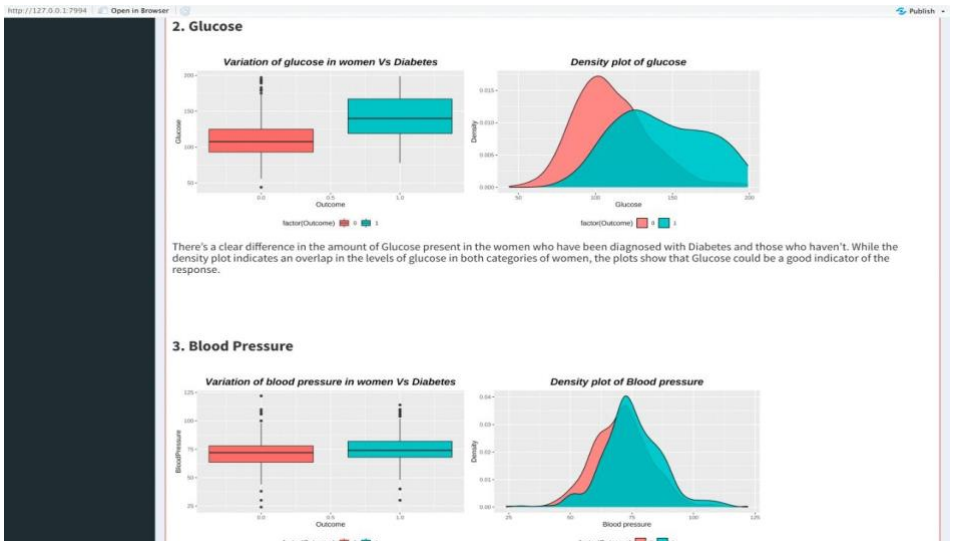
Outcome is the response variable which is a binary classifier with values 0 and 1 indicating non-diabetic and diabetic respectively.

DATA PREPROCESSING

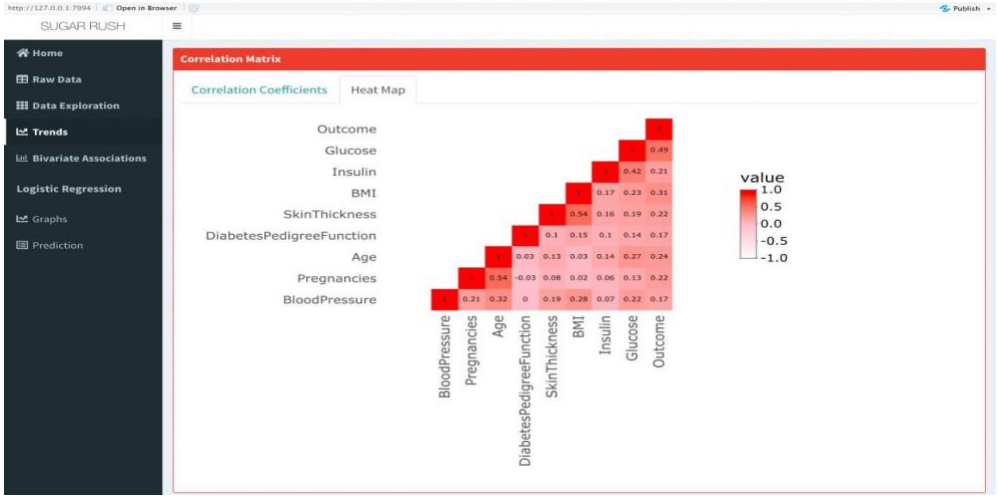
In the first glance, dataset appeared to be clean. On deeper analysis, the dataset revealed abnormalities. Variables like Glucose, Skin thickness, Blood pressure, BMI and Insulin had few records with value 0. It is a fact that these variables cannot have 0 value. Hence it was replaced using kNN imputation.

EXPLORATORY DATA ANALYSIS

We have performed analysis on each variable in the dataset using Box plots and Density plots to determine how good or bad the variable is for prediction.

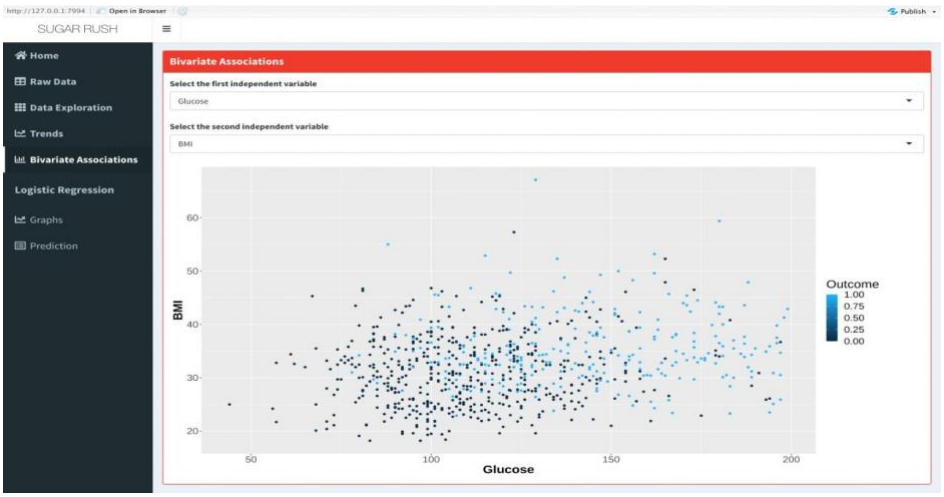


After analysis, we came to the conclusion that due to less overlapping and significant difference between levels of boxplot, Glucose is a good indicator of the response when compared to the other variables.



The heat map depicts the correlation values between each variable which is depicted by varying intensity of the colour. From the heat map, we can conclude that the highest correlation is between Skin Thickness & BMI and Age & Pregnancy with correlation value 0.54 each.

Glucose has the highest correlation with the response variable (Outcome) with correlation value 0.49.

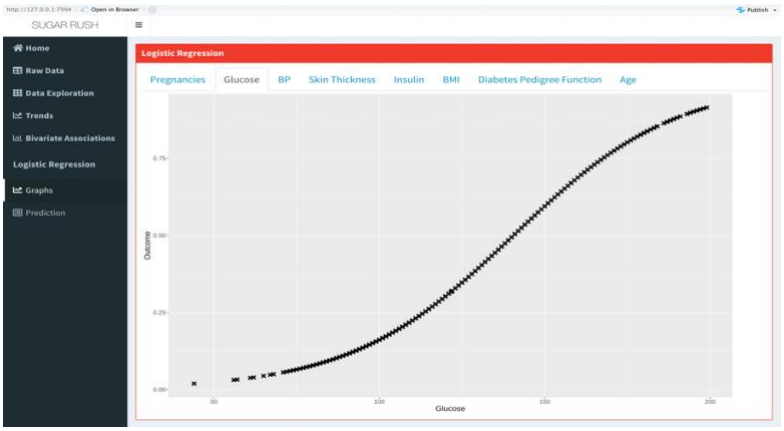


The grouped scatter plots determine whether the relationship between two predictor variables differs between the two groups : diabetic and non-diabetic.

The dashboard provides a grouped scatter plot for each combination of predictor variables depending on the user inputs.

LOGISTIC REGRESSION MODEL

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.



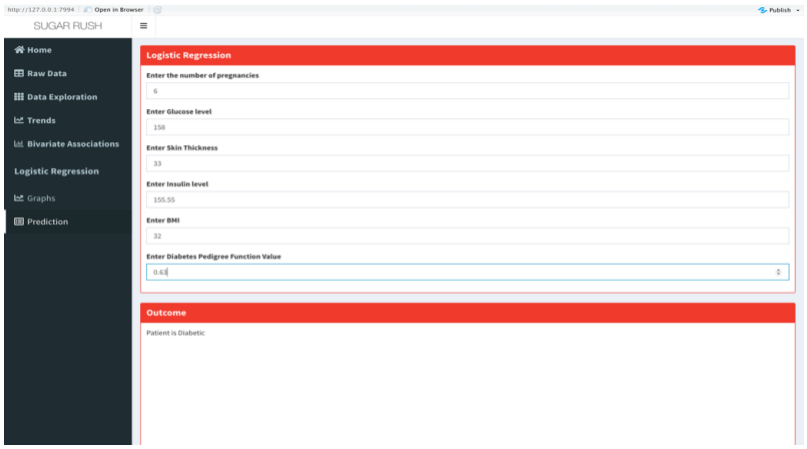
For the model we consider only those variables whose effect on prediction is significant. Based on the p-values, we have eliminated Age and Blood pressure from the model.

Our final model is given as:

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

PREDICTION

Dashboard also included a tab for users to check if they are diabetic or not by entering the values of Number of pregnancies, Glucose, Skin Thickness, Insulin, BMI and DPF.



CONCLUSION AND FUTURE WORK

The Pima Women's dataset was analyzed and explored in detail to conclude that Glucose has the highest correlation with outcome. 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%.

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.

LINKS

- Link to the Presentation: https://drive.google.com/file/d/1vETj0IVRzW4qvi342Lu7SYYDqWJUHlGS/view?usp=sharing
- Link to the Dashboard Screenrecording: https://drive.google.com/file/d/1mGnYDJZ0fsv7SDYePe6yHCWxdUKFWXDU/view?usp=sharing

REFERENCES

- www.kaggle.com
- https://rpubs.com/soodrk/578110
- https://shiny.rstudio.com/