Diabetes Prediction

Mohd. Rizwan Shaikh
IMT2019513

October 3, 2021

Abstract

The aim of this machine learning assignment is to predict whether a person has diabetes or not based on certain features. For this, we implement Logistic Regression using Gradient Descent and Newton's Method of Optimization. We also implement Naive Bayes for Univariate Gaussian. At the end, we also explore the possibility of implementing a better performing multivariate Gaussian.

Contents

1 Introduction

Diabetes is a chronic (long-lasting) health condition that affects how a person's body turns food into energy. Most of the food that a person eats is broken down into sugar and released into bloodstream. When blood sugar goes up, it signals the pancreas to release insulin. Insulin acts like a key to let the blood sugar into the body's cells for use as energy.

If a person has diabetes, his body either doesn't make enough insulin or can't 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 the bloodstream. Over time, that can cause serious health problems, such as heart disease, vision loss, and kidney disease.

Through this machine learning assignment, we try to predict whether a person has diabetes or not. We look at features such as number of pregnancies, glucose, blood pressure, skin thickness, insulin, BMI and age and try to generate results using the same. For prediction, we implement logistic regression and Naive Bayes for Univariate Gaussian. We also explore the possibility of implementing a better performing multivariate Gaussian.

2 Initial Data Processing

2.1 Null values

In the given data set, there are no null values. But if we look closely, we realize that zeros are used as placeholders for null values. Various features have zero values which is practically impossible. So we remove these zeros to clean our data.

2.2 Removing zeros

Table 1: Number of zeros in each feature

Feature	Number of zeros
Pregnancies	111
Glucose	5
BloodPressure	35
SkinThickness	227
Insulin	374
BMI	11
DiabetesPedigreeFunction	0
Age	0

Zero as entry does not make sense in these features (expect pregnancies). We assume these values as illogical values and remove the same.

We drop the rows where Glucose, BloodPressure and BMI are zero because of very few number of zeros in them.

We drop SkinThickness and Insulin since these features have large number of zeros as entries in them.

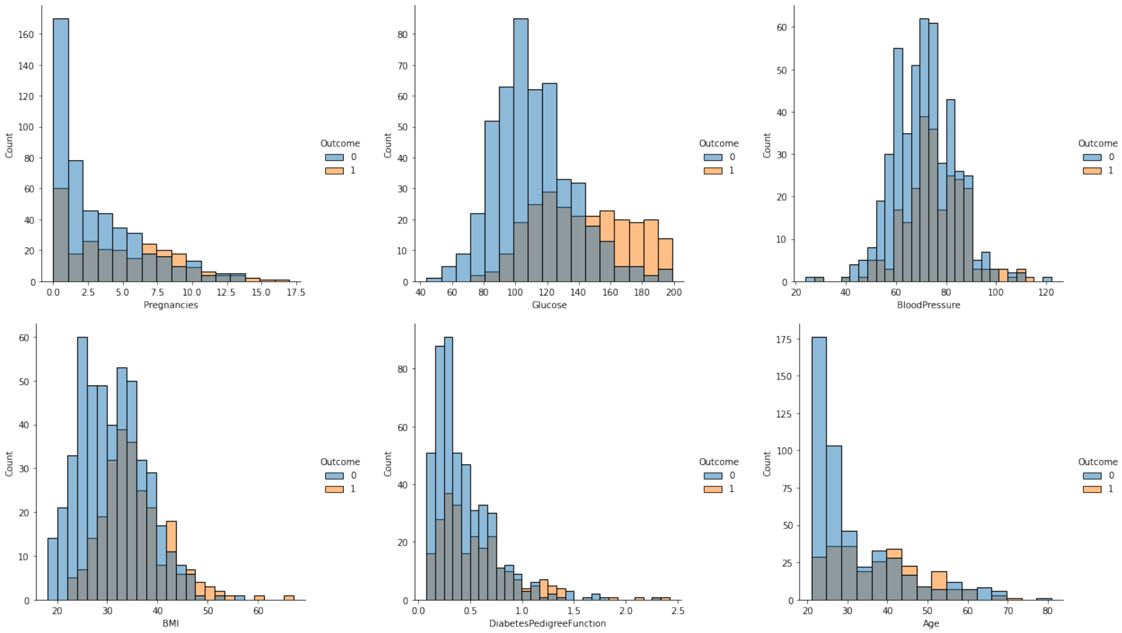
2.3 Other Illogical Values

There are no '?' as placeholders in the data set. Also, there are no duplicate rows.

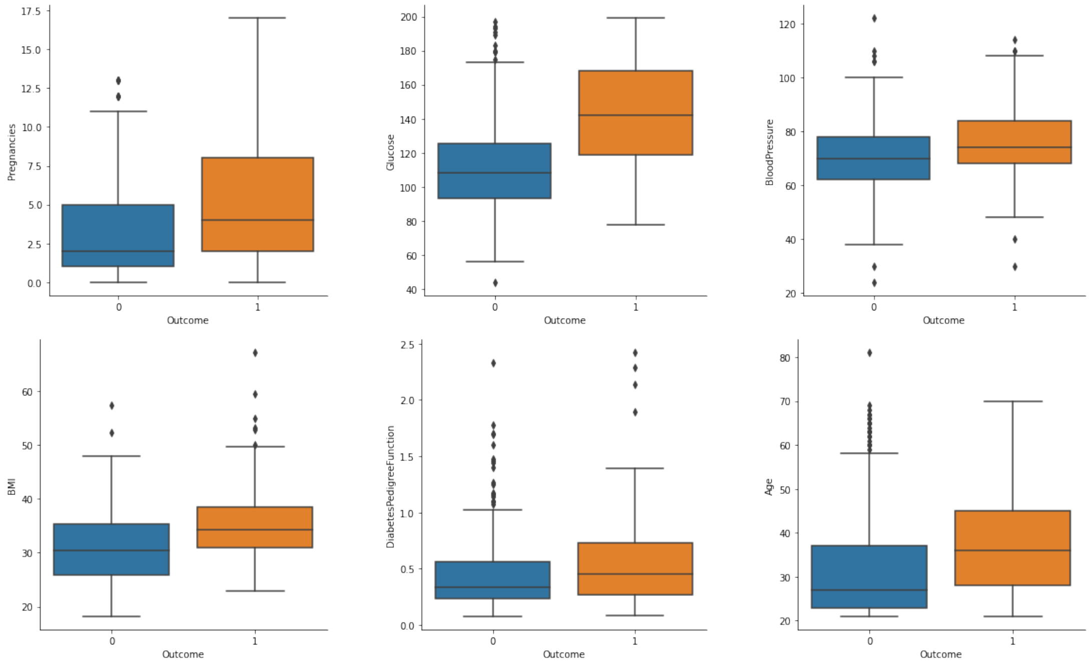
3 Exploratory Data Analysis and Outlier Removal

3.1 Data Visualization and Outliers

The histograms of the features after removal of illogical values is as follows:



The following boxplots gives a rough estimate of the outliers in each feature:



3.2 Outlier Removal and skew removal

The number of outliers in each feature is shown below:

Table 2: Number of outlier values in each feature

Feature	Number of outlier values
Pregnancies	10
Glucose	11
BloodPressure	11
BMI	9
DiabetesPedigreeFunction	18
Age	31

For all these features, we remove the rows containing the outlier values since these are few in number. We also remove the skew in Pregnancies, DiabetesPedigreeFunction and Age histogram. For Pregnancies and DiabetesPedigreeFunction, we apply the following transformation:

$$x \leftarrow np.log(x + 1)$$

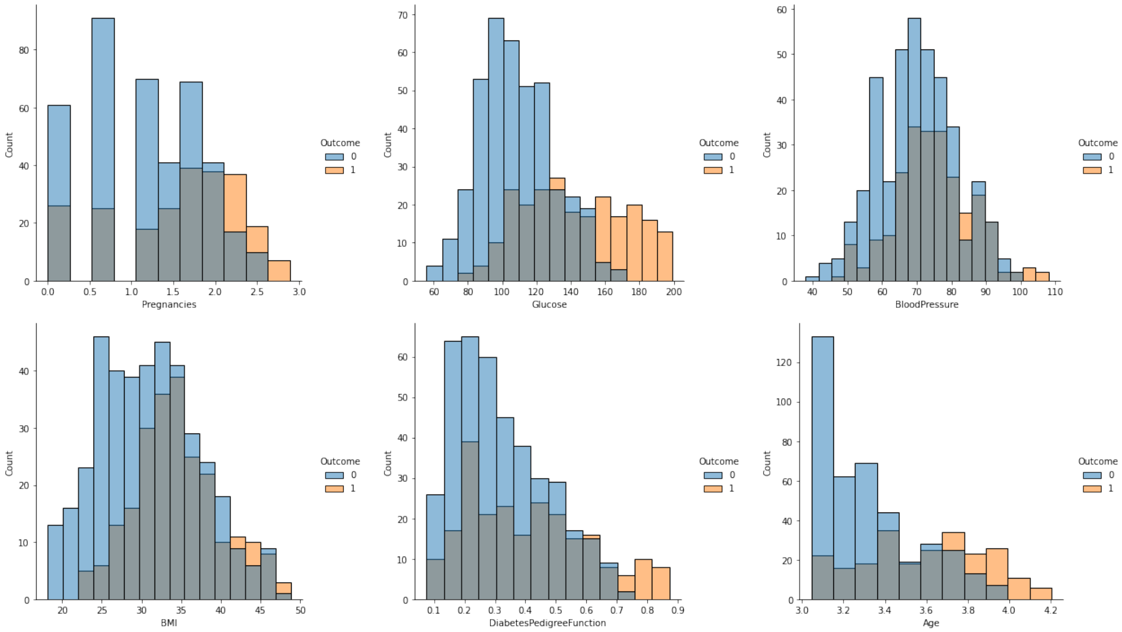
Whereas, for Age, we apply the following transformation:

$$x \leftarrow np.log(x)$$

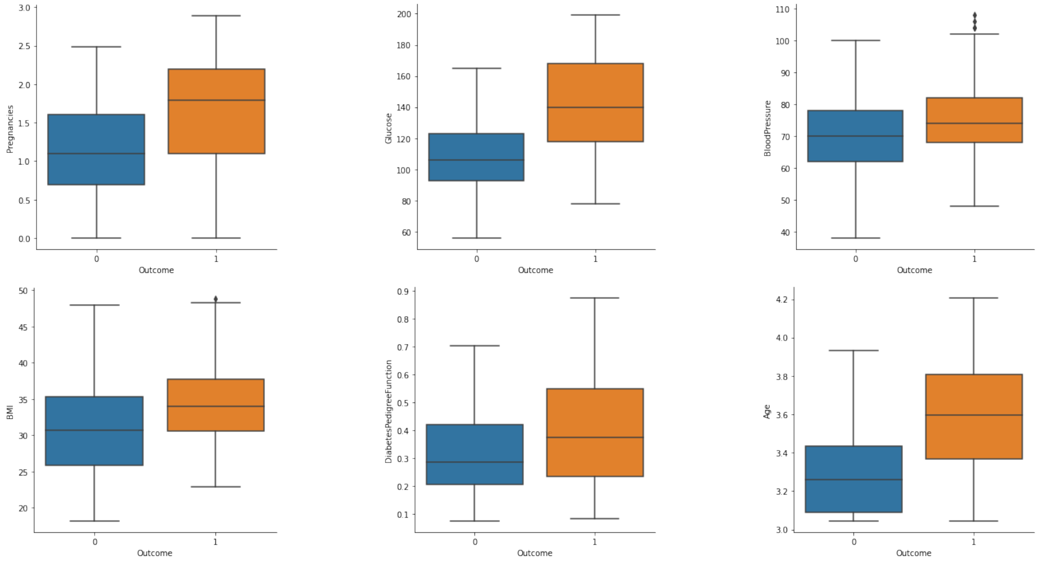
3.3 Processed data

After removal of outlier values and skews, we get the final data which can be used to implement the models.

The histograms of the features after cleaning the data is as follows:



The boxplots of the features are as follows:



We use 70% of this data for training our model. Before training, we normalize the data and add a column of ones to the feature matrix to calculate the constant term where ever necessary.

4 Logistic Regression

We consider all the features to implement multivariate logistic regression. To implement logistic regression, we use gradient descent method and newton's method of optimization.

4.1 Gradient Descent Method

For this method, the learning rate is considered as 0.5 and the number of iterations is 2500.

Roughly, the model gives an accuracy of 80% to 85%

The output of the model on testing it on both training and testing data set is shown below:

```
Weight Matrix: [−6.83547122 −0.21640168  6.34033716 −1.11474954  2.59515928
3.16694676  4.35613822]

On Testing Dataset:
Hit count: 157
Miss count:33
Accuracy: 82.63157894736842%

On Training Dataset:
Hit count: 361
Miss count:83
Accuracy: 81.30630630630631%
```

The loss curve for the same is as follows:

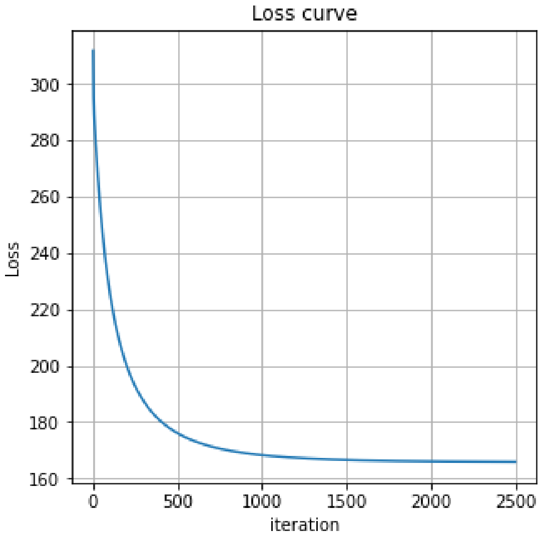


Figure 5: Loss vs Iterations for Gradient Descent

4.2 Newton's Method of Optimization

For this method, the number of iterations is 10.

Roughly, the model gives an accuracy of 80% to 85%

The output of the model on testing it on both training and testing data set is shown below:

```
Weight Matrix: [-7.09705727 -0.23171453  6.61154119 -1.23195123  2.74498098
3.32119871  4.52119217]

On Testing Dataset:
Hit count: 156
Miss count:34
Accuracy: 82.10526315789474%

On Training Dataset:
Hit count: 360
Miss count:84
Accuracy: 81.08108108108108%
```

The loss curve for the same is as follows:

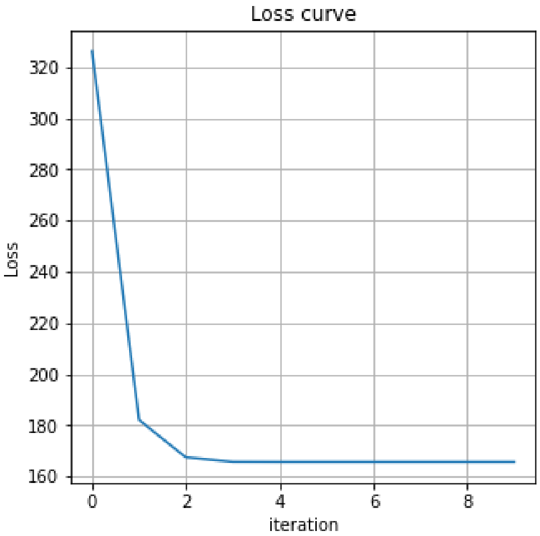


Figure 6: Loss vs Iterations for Newton's Method

5 Naive Bayes(Univariate Gaussian)

The outcome of Naive Bayes Model for Univariate Gaussian for all features are shown below:

5.1 Pregnancies

```
On Testing Dataset:
 Hit count: 121
 Miss count:69
 Accuracy: 63.68421052631579%

On Training Dataset:
 Hit count: 285
 Miss count:159
 Accuracy: 64.1891891891892%
```

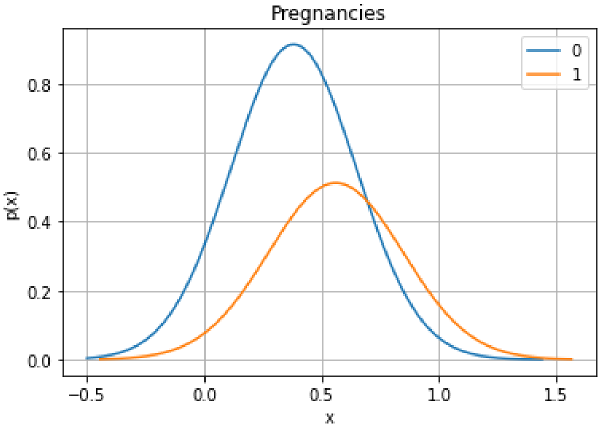


Figure 7

5.2 Glucose

```
On Testing Dataset:
 Hit count: 135
 Miss count:55
 Accuracy: 71.05263157894737%

On Training Dataset:
 Hit count: 320
 Miss count:124
```

```
Accuracy: 72.07207207207207%
```

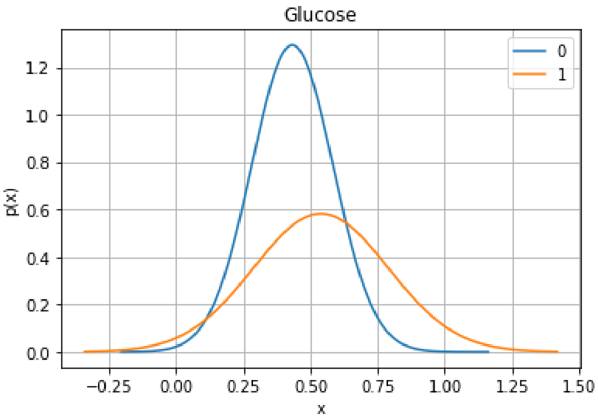


Figure 8

5.3 BloodPressure

```
On Testing Dataset:
Hit count: 105
Miss count:85
Accuracy: 55.26315789473684%

On Training Dataset:
Hit count: 272
Miss count:172
Accuracy: 61.26126126126126%
```

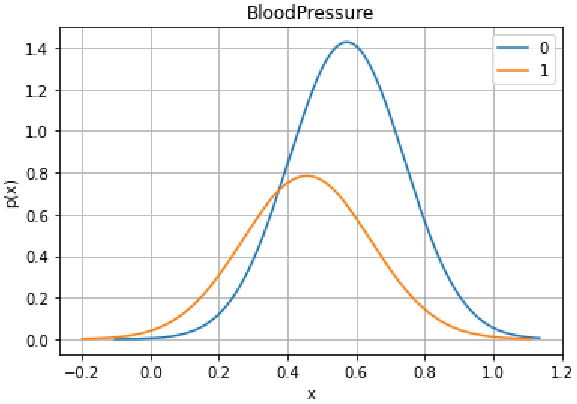


Figure 9

5.4 BMI

```
On Testing Dataset:
Hit count: 119
Miss count:71
Accuracy: 62.63157894736842%

On Training Dataset:
Hit count: 281
Miss count:163
Accuracy: 63.288288288288285%
```

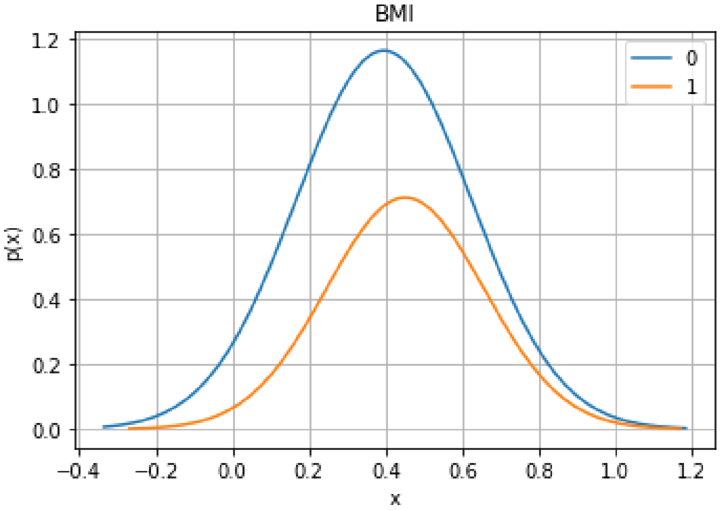


Figure 10

5.5 DiabetesPedigreeFunction

```
On Testing Dataset:
Hit count: 124
Miss count:66
Accuracy: 65.26315789473684%

On Training Dataset:
Hit count: 297
Miss count:147
Accuracy: 66.89189189189189%
```

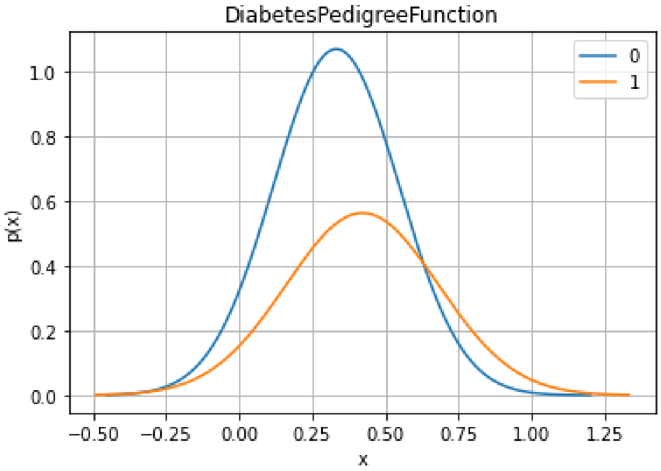


Figure 11

5.6 Age

```
On Testing Dataset:
 Hit count: 134
 Miss count:56
 Accuracy: 70.52631578947368%

On Training Dataset:
 Hit count: 321
 Miss count:123
 Accuracy: 72.29729729729729%
```

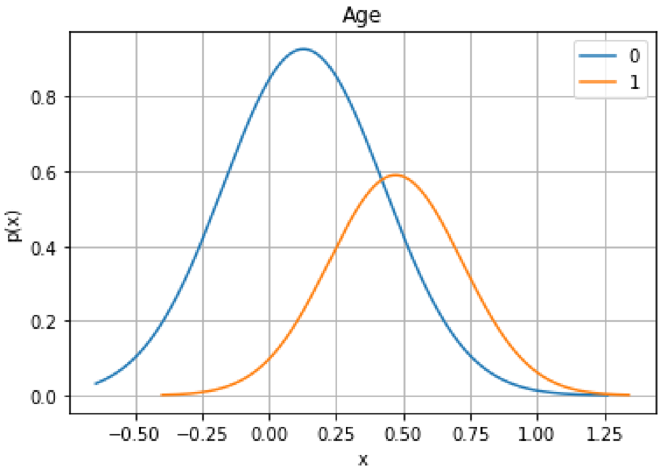


Figure 12

6 Naive Bayes(Multivariate Gaussian)

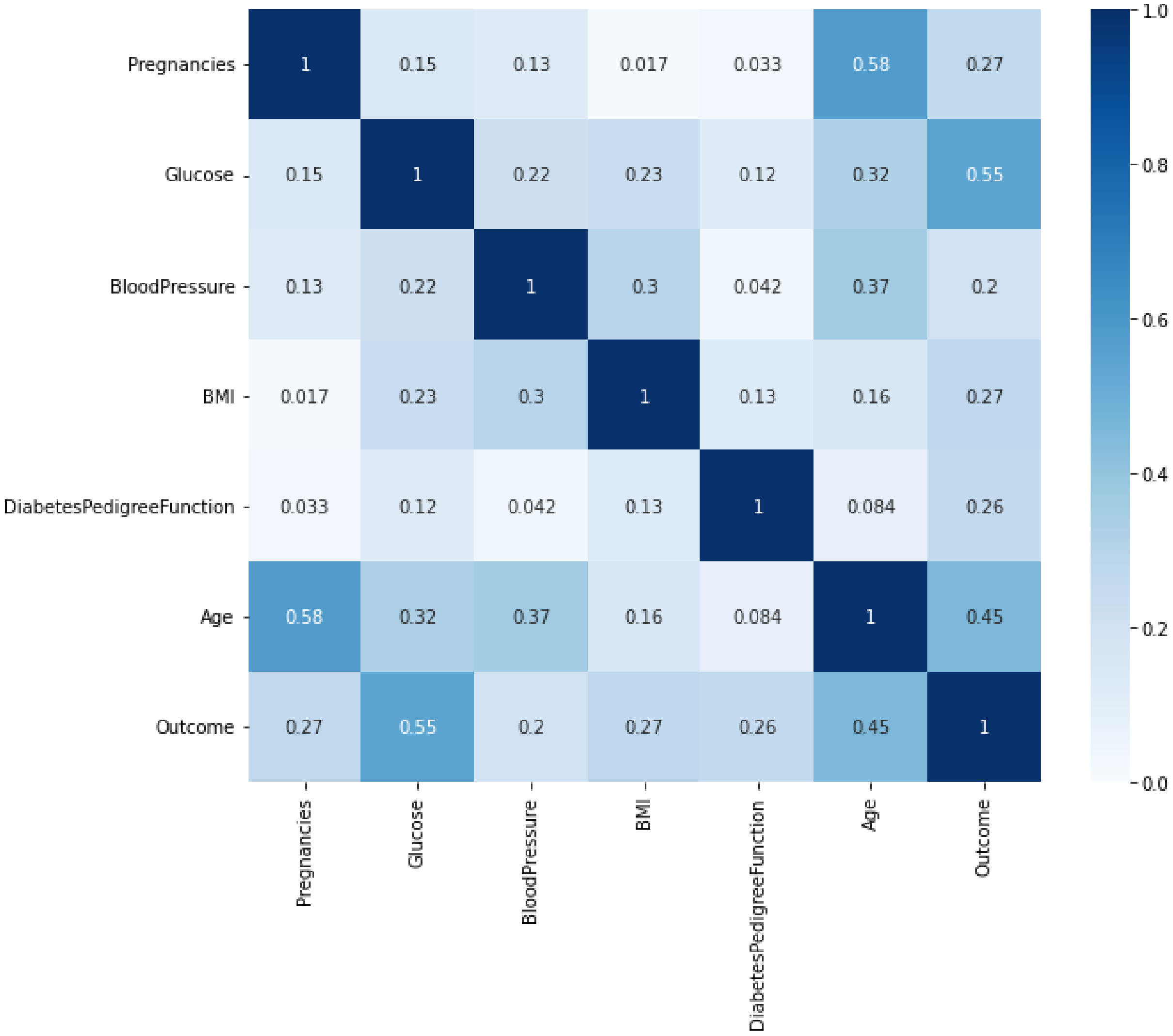


Figure 13: Correlation Heatmap

The above heat-map gives us a correlation matrix where each cell denotes the correlation coefficient of the corresponding features. We interpret correlation coefficient as follows:

- A positive correlation coefficient means both the features have a positive relationship. It implies that both the features will increase together and decrease together.
- Similarly, a negative correlation coefficient implies that the features have negative relationship. It implies that one feature will decrease as other increases.
- A correlation coefficient of zero implies that both the features are independent of each other.

While implementing multivariate gaussian, we assume the features to be independent of each other. For this reason, we would prefer features which are strongly independent of each other. Thus, features whose correlation coefficients are closer to zero would be preferred over the features whose coefficients are farther away from zero.

Some pair of features that satisfy the above condition are as follows:

- BMI and Pregnancies (Correlation Coefficient: 0.017)
- DiabetesPedigreeFunction and Pregnancies (Correlation Coefficient: 0.033)
- DiabetesPedigreeFunction and BloodPressure (Correlation Coefficient: 0.042)
- Age and DiabetesPedigreeFunction (Correlation Coefficient: 0.084)

We would prefer these set of features for multivariate gaussian because of their strong linear independence as compaired to other pairs of features.