EVALUATION AND PREDICTION OF DIABETES IN INDIA

A project work done in partial fulfillment of the "Certificate course on Data Analytics & Business Intelligence"

Submitted by :

Abhilasha Chatterjee

Certificate course on Data Analytics & Business Intelligence Batch-09

Shaheed Sukhdev College of Business Studies

December 2023

DECLARATION

I, Abhilasha Chatterjee, declare that this project titled “Evaluation and Prediction of Diabetes in India” is the original work done by me under the guidance of Dr. Rishi Ranjan Sahay, Assistant Professor, Shaheed Sukhdev College of Business Studies, University of Delhi.

I further declare that this work is for my certificate course in Data Analytics and Business Intelligence.

Name : Abhilasha Chatterjee

ACKNOWLEDGEMENT

I would like to express my special thanks of gratitude to our teacher Dr. Rishi Ranjan Sahay, who gave us the golden opportunity to do this wonderful project. This project helped me in doing a lot of research and discovering many new things.

I am overwhelmed in all humbleness and gratefulness to acknowledge my learning to all those who have helped me to put these ideas, well above the level of simplicity and into something concrete.

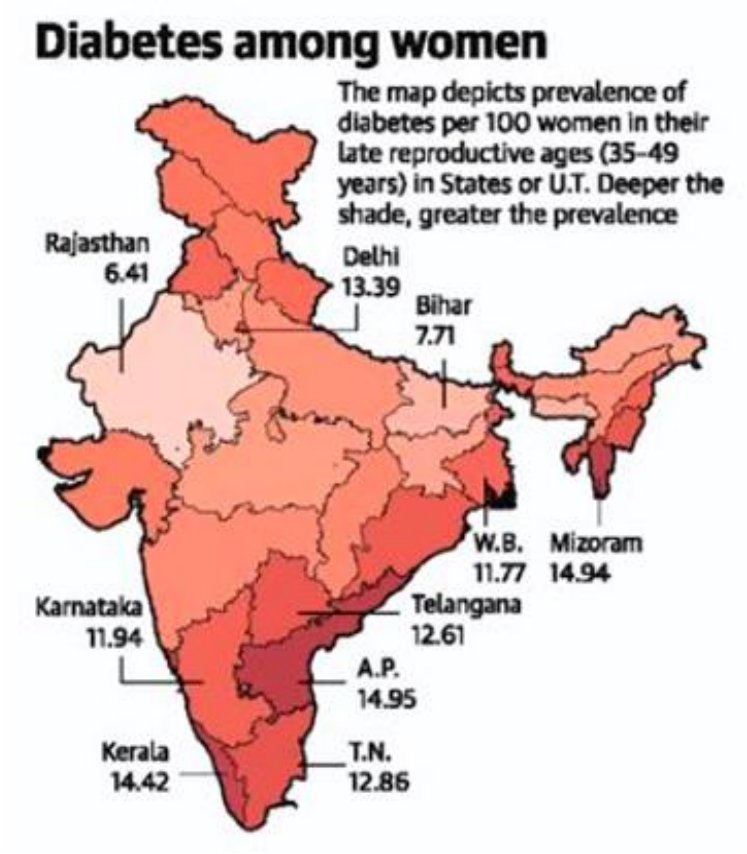
Thanking You

Abhilasha Chatterjee

ABSTRACT

This research study is focused on the analysis of diabetes in female dataset and how it will perform if I want to do a predictive analysis using machine learning algorithm. This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset. For analysis of the given dataset I have applied the classification model of Logistic Regression. I have done a performance measurement on the basis of that.

INTRODUCTION



Diabetes is a major public health concern in India, as it can lead to several serious health complications associated with metabolic disorder which include cardiovascular disease, kidney disease, blindness and amputations as well. It is also a leading cause of death in India.
Diabetes is a chronic disease affecting millions of people worldwide. In India, there is an expectation of significant increase in diabetes, which may reach to 123 million by 2040.

Prevention and management of diabetes is critical to reduce the burden of this disease in India. Lifestyle changes like regular exercise and a healthy diet can help prevent or delay the onset of diabetes.
Management of diabetes involves controlling blood sugar level through a combination of lifestyle changes and medications. Regular monitoring of sugar levels, regular check-ups with healthcare provider and education on how to manage it are also important. Screening programs can also help identify people with diabetes early, so they can receive appropriate treatment.

The use of classification is to extract the model and to describe the classes. Classification predicts categorical, labels, models continuous valued functions. Classification organizes and categorizes data in distinct classes.
The focus here is to apply the classification algorithm for the given dataset i.e. Logistic Regression Model.

According to the World Health Organization (WHO), 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.
There are three main types of diabetes mellitus. Type 1 and type 2 diabetes both occur when the body cannot properly store and use glucose, which is primary energy source. Sugar, or glucose, collects in the blood and does not reach the cells that need it, which can lead to serious complications.

Type 1 diabetes usually begins in childhood or early adulthood; but its onset can occur in adults. In type 1 diabetes, a person with a genetic predisposition who is exposed to a precipitating event, such as a viral infection, experiences autoimmune destruction of the beta cells.
Type 2 diabetes is more likely to appear as people age, but many children are now starting to develop it. In this type, the pancreas produces insulin, but the body cannot use it effectively.
There has been a substantial increase in the number of cases of type 2 diabetes diagnosed in young children

RESEARCH OBJECTIVES

- To apply a regression model to measure the amount of diabetic female patients in India.
- To develop my own prediction model by selecting variables from the array of indicators categorized under different heads.

The two models' accuracy of prediction is then compared against each other to address the differences in the model.

METHODOLOGY

This project covers both theoretical and speculative analysis of the problem of diabetes as a major concern. Along with the theory, the speculation includes using correlation and logistic regression.

DATASET

This dataset is from the National Institute of Diabetes and Digestive and Kidney Diseases.

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6.0	148.0	72.0	35.0	0.0	33.6	0.627	50.0	1.0
1	1.0	85.0	66.0	29.0	0.0	26.6	0.351	31.0	0.0
2	8.0	183.0	64.0	0.0	0.0	23.3	0.672	32.0	1.0
3	1.0	89.0	66.0	23.0	94.0	28.1	0.167	21.0	0.0
4	0.0	137.0	40.0	35.0	168.0	43.1	2.288	33.0	1.0
5	5.0	116.0	74.0	0.0	0.0	25.6	0.201	30.0	0.0
6	3.0	78.0	50.0	32.0	88.0	31.0	0.248	26.0	1.0
7	10.0	115.0	0.0	0.0	0.0	35.3	0.134	29.0	0.0

Shape of the dataset

```
(771, 9)
```

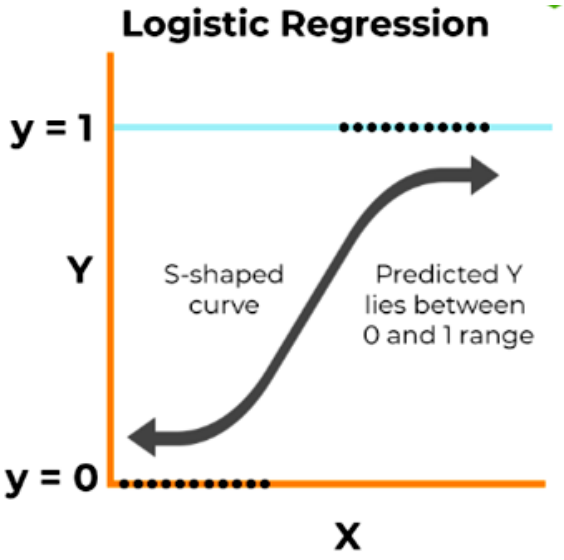
Extracted Information

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 771 entries, 0 to 770
Data columns (total 9 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Pregnancies               768 non-null    float64
 1   Glucose                   768 non-null    float64
 2   BloodPressure             770 non-null    float64
 3   SkinThickness             768 non-null    float64
 4   Insulin                   768 non-null    float64
 5   BMI                       770 non-null    float64
 6   DiabetesPedigreeFunction  768 non-null    float64
 7   Age                       770 non-null    float64
 8   Outcome                   768 non-null    float64
dtypes: float64(9)
memory usage: 54.3 KB
```

Splitting the dataset into training and testing data

```
In [16]: from sklearn.model_selection import train_test_split
         x_train, x_test, y_train, y_test = train_test_split(x, y, test_size = 0.2)

In [17]: x_train.shape

Out[17]: (616, 8)

In [18]: y_train.shape

Out[18]: (616,)

In [19]: x_test.shape

Out[19]: (155, 8)

In [20]: y_test.shape

Out[20]: (155,)
```

DEVELOPMENT OF LOGISTIC REGRESSION MODEL



LOGISTIC REGRESSION

Logistic regression analysis is conducted when the dependent variable is dichotomous (binary). It is used in predictive analysis like other regression models. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval, or ratio-level independent variables.

A logistic regression analysis is undertaken when we have a binary dependent variable in our model. Moreover, the effects of outliers are minimized in logistic regression. Linear regression models are highly affected by outliers as the best fit line shifts to minimize the distance between the predicted value and actual value.

Assumptions of the model :

1. The outcome is binary
   - The outcome of a logistic regression model is binary, i.e., can only take two values like bankrupt and non-bankrupt.
2. The logistic of the outcome and independent variable have a linear relationship
   - The outcome of the model and the explanatory variable should have a linear relationship.
3. No homoskedasticity
   - The logistic model doesn't require an assumption of homoskedasticity, i.e., when the error term (that is, the "noise" or random disturbance in the relationship between the independent variables and the dependent variable) is the same across all values of the independent variables.
4. No severe multicollinearity
   - The logistic regression model necessitates that there should not be high multi-collinearity between the independent variables.

Function of the model

$$f(x) = \frac{L}{1 + e^{-k(x-x_0)}}$$

$f(x)$ = output of the function
L = the curve's maximum value
k = logistic growth rate or steepness of the curve
x_0 = the x value of the sigmoid midpoint
x = real number

CLASSIFICATION

Classification is a technique for identifying and grouping data in such a way that based on a value of the target attribute, the entire dataset can be qualified to belong to a class. This is one of the techniques used in data mining to identify the data behavior patterns.

Classification is two-step process:

1. Learning or training step where data is analyzed by a classification algorithm.
2. Testing step where data is used for classification and to estimate the accuracy of the classification

Logistic Regression is the most common method used for binary classification problems.
The function used in Logistic regression is the logistic function which is an S-shaped curve that can take any number and map it into a value between 0 and 1.
Input values are combined linearly using coefficient values to predict an output value.

As we are using logistic function for the classification problem, we won't use the probabilities directly.
We will convert the probabilities to a binary class value, like:

- If probability < 0.5 assign : 0
- If probability >= 0.5 assign : 1

METHODOLOGY

This research dataset is divided into two parts, two-thirds of the data is used as a training set, and one-third of the dataset is defined as a testing set to evaluate the performance of several classifiers. All classifiers were fitted to the same training and testing data.

Independent Variables for the dataset

1.	**Pregnancies**	Number of times pregnant
2.	**Glucose**	Glucose concentration a 2 hours in an oral glucose tolerance test
3.	**BloodPressure**	Diastolic blood pressure (mm Hg)
4.	**SkinThickness**	Triceps skin fold thickness (mm)
5.	**Insulin**	2-Hour serum insulin (mu U/ml)
6.	**BMI**	Body mass index (weight in kg/(height in m)^2)
7.	**DiabetesPedigreeFunction**	Diabetes pedigree function
8.	**Age**	Age (years)

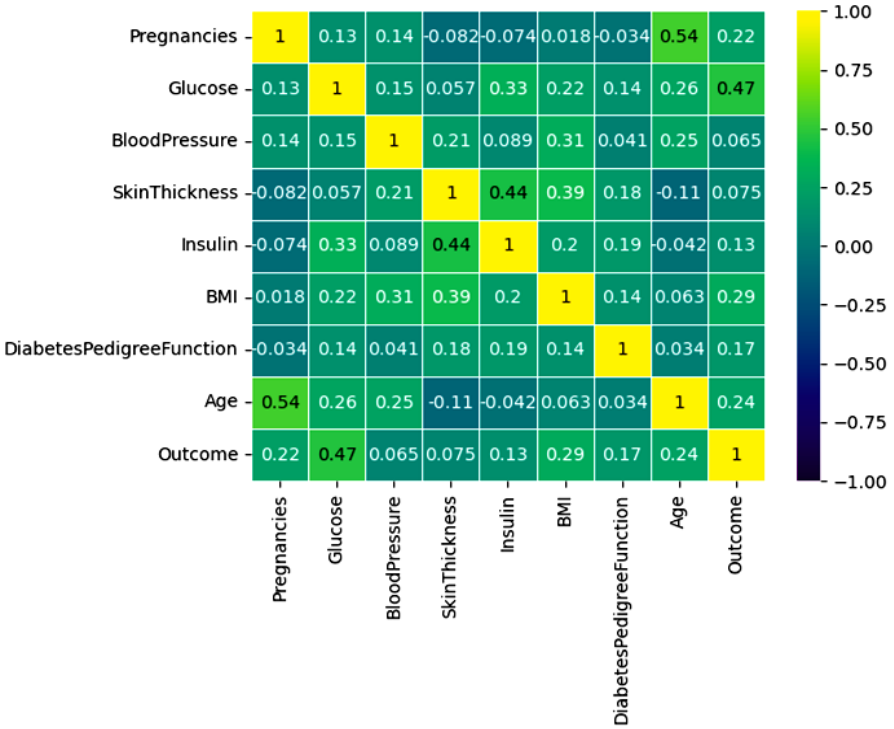
The dataset consists of 8 medical independent variables [pregnancies, glucose, blood pressure, skin thickness, insulin, BMI, diabetes pedigree function and age] and one target dependent variable (outcome).

"Outcome" is a binary target variable that has a value of 1 for diabetes and 0 for no diabetes. 268 of the 768 are 1, the others are 0.

All records for patients in the dataset are females greater than 21 years old of Pima Indian heritage.

CORRELATION OF THE VARIABLES

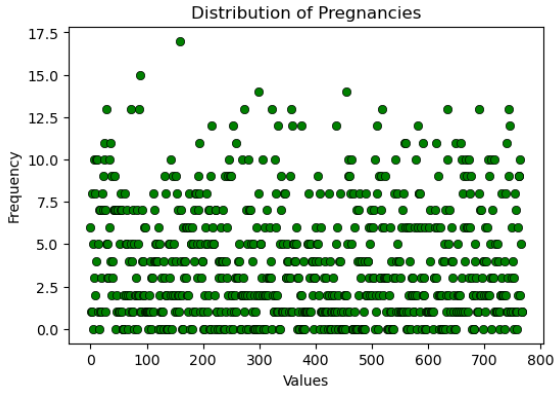
To analyse the correlation between the dependent variables and status we use Karl Pearson correlation coefficient.

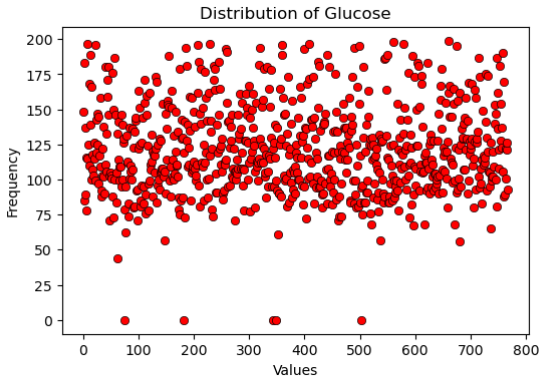


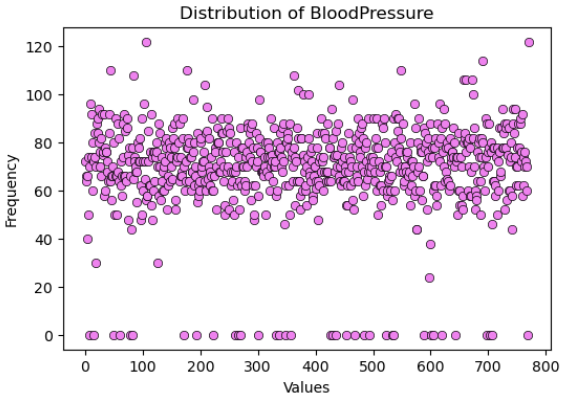
Observations

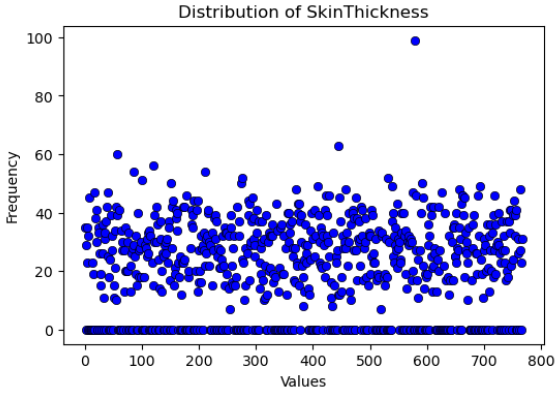
1. We observe that Glucose and Outcome have a high correlation as Glucose seems to be the most important feature in model training.
2. We can also see a low to moderate degree of correlation, i.e., 0 to 0.5 between the dependent variables. BMI, Pregnancies and Age are also important features.
3. There is also negative correlation between skin thickness and age.

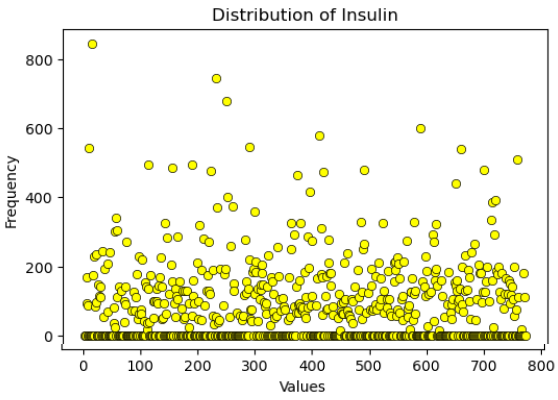
DISTRIBUTION OF THE VARIABLES

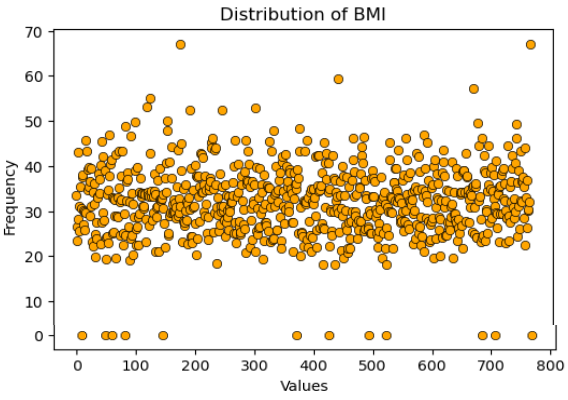


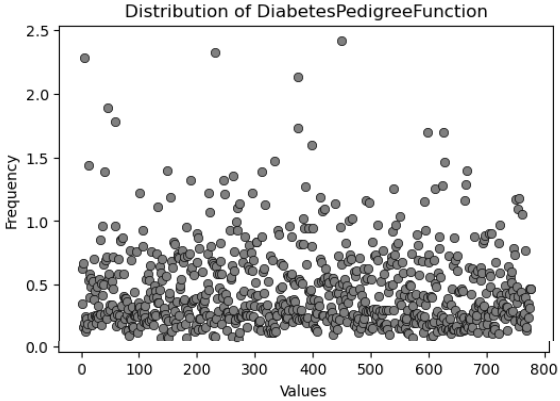


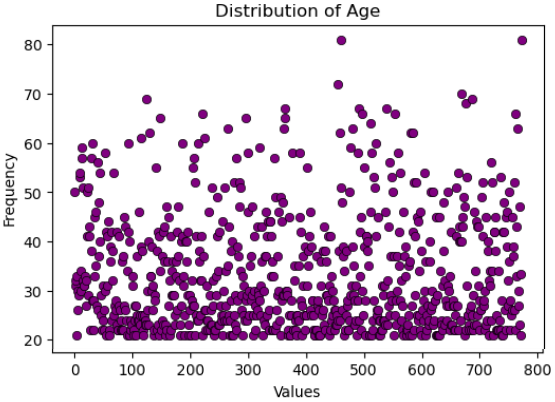






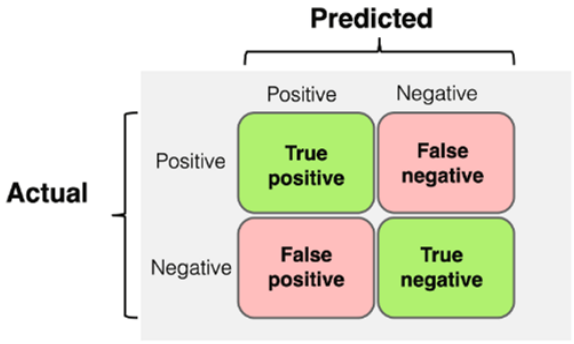






RESULTS OF CLASSIFICATION ALGORITHM

To see if the model gives good results at identifying 1's and 0's, we check the confusion matrix. A confusion matrix is a table that is used to define the performance of a classification algorithm by visualizing and summarizing it.

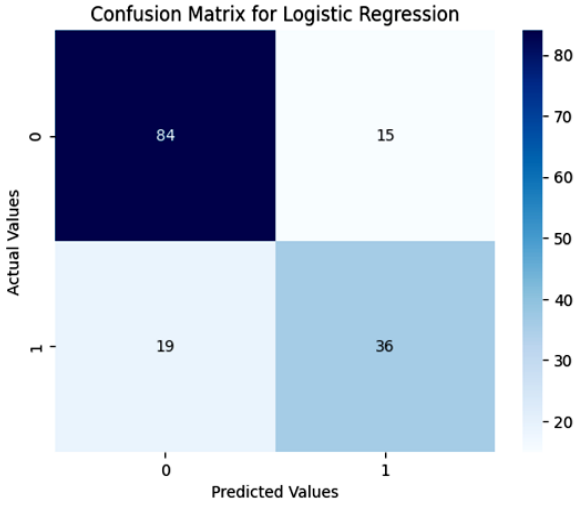


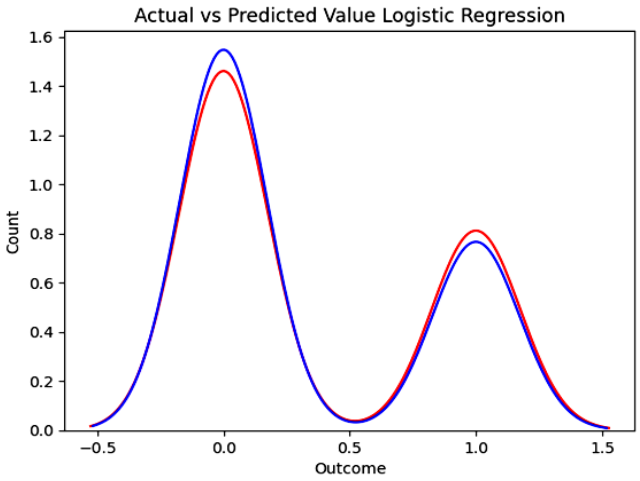
The results of diabetes dataset for Logistic Regression Model classifier is shown :

True Negative: 445	False Negative: 55
False Positive: 112	True Positive: 156

- Logistic Regression Model

```
Training Accuracy of Logistic Regression Model is 0.805940594059406
Test Accuracy of Logistic Regression Model is 0.7751479289940828
```



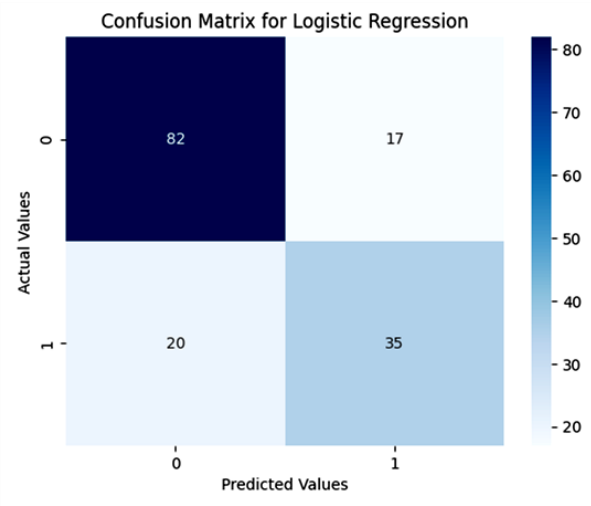


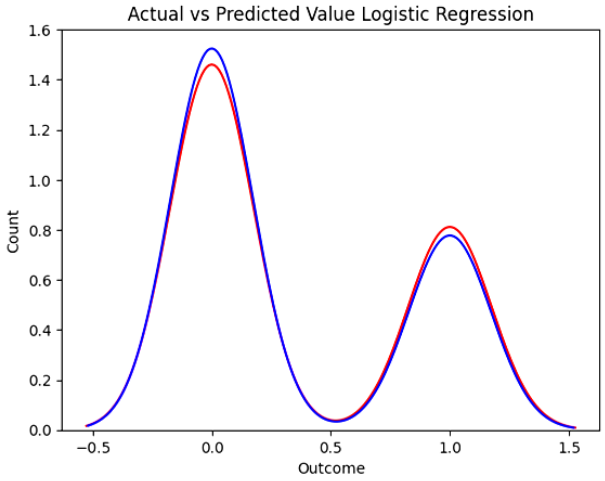
➢ Classification report

	precision	recall	f1-score	support
0	0.80	0.91	0.85	117
1	0.69	0.48	0.57	52
accuracy			0.78	169
macro avg	0.75	0.69	0.71	169
weighted avg	0.77	0.78	0.76	169

- K-nearest Neighbors

```
Training Accuracy of KNN Model is 0.8376237623762376
Test Accuracy of KNN Model is 0.7514792899408284
```



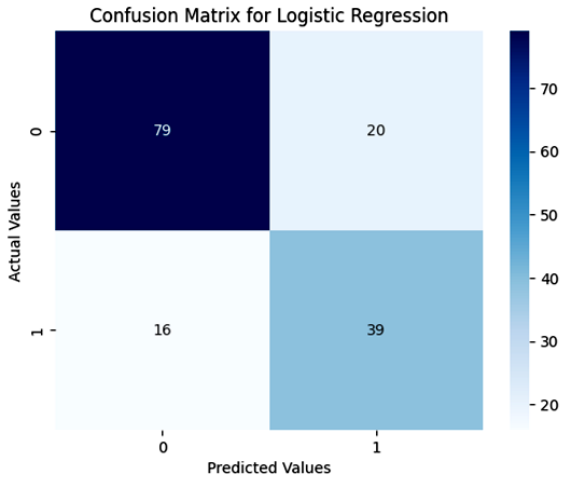


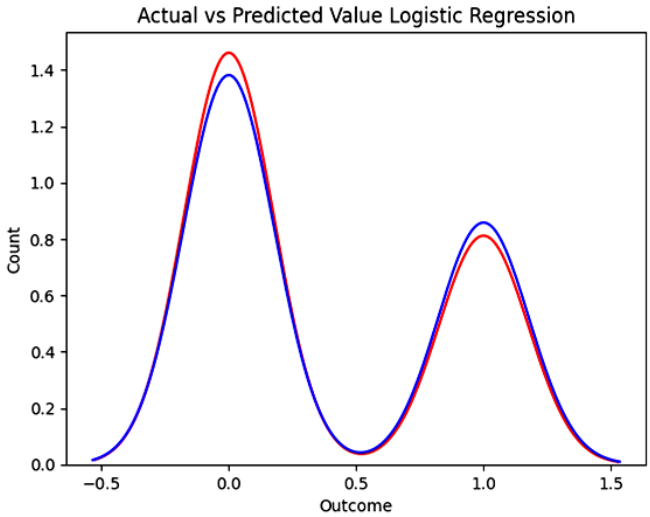
➢ Classification report

	precision	recall	f1-score	support
0	0.79	0.88	0.83	117
1	0.63	0.46	0.53	52
accuracy			0.75	169
macro avg	0.71	0.67	0.68	169
weighted avg	0.74	0.75	0.74	169

- Random Forest

```
Training Accuracy of Random Forest Model is 0.80990099009901
Test Accuracy of Random Forest Model is 0.7692307692307693
```





➢ Classification report

	precision	recall	f1-score	support
0	0.78	0.92	0.85	117
1	0.71	0.42	0.53	52
accuracy			0.77	169
macro avg	0.75	0.67	0.69	169
weighted avg	0.76	0.77	0.75	169

The Logistic Regression model has an accuracy of 78%.
The K-nearest Neighbors model has an accuracy of 75%.
The Random Forest model has an accuracy of 77%.

CONCLUSION

From the exploratory data analysis, this can be concluded that the risk of diabetes depends upon the following factors:

1. Glucose level
2. Number of pregnancies
3. Skin Thickness
4. Insulin level
5. BMI

With an increase in Glucose level, insulin level, BMI and number of pregnancies, the risk of diabetes increases. However, the number of pregnancies have a strange effect on the risk of diabetes which can't be explained by the data. The risk of diabetes also increases with increase in skin thickness.

According to the classification models, Logistic Regression is the best algorithm as it outperformed Random Forest and KNN models with 78% accuracy. The accuracy of the model can be improved by increasing the size of the dataset.
The dataset used here is very small and had only 768 rows.

BIBLIOGRAPHY

(2017) Pima Indians Diabetes Database
Link : https://www.kaggle.com/uciml/pima-indians-diabetes-database