**Linear Regression using python on diabetes dataset**

**Data Set Description:**

In this project, we will build a simple but robust linear regression model from scratch in Python and use it to predict the Blood Sugar of Diabetes Patients from their BMI and BP data.

Read more about Body Mass Index and how it impacts people’s health in various ways, and ultimately how it affects Blood Sugar among diabetes patients Here: <https://en.wikipedia.org/wiki/Body_mass_index>.

Data was taken from  <http://www4.stat.ncsu.edu/~boos/var.select/diabetes.tab.txt> is showing Blood Sugar (in mg/l) of 442 diabetes patient with 10 baseline variables, age, sex, body mass index, average blood pressure, six blood serum measurements and finally target(Blood Sugar of patient).

**Agenda of the Project:**

Fitting data around straight line equation y = mx+b

Computing errors via Mean Square Error.

Minimizing Errors using partial derivatives aka Gradient Descent

Computing new m & b by using Gradient Descent with Learning Rate over many iterations

Use this new m & b to predict Blood Sugar of sample/test data of BMI and BP.

**Data Preprocessing:**

For the DataSet which I have luckily has all integer values which makes clean of data easy😊.

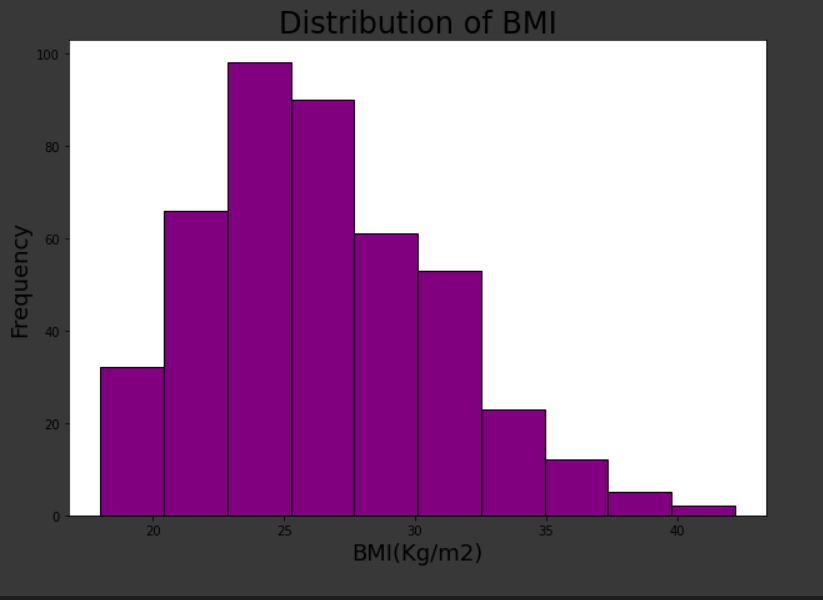
Selection of features:

1. As we have seen from the above, data has 9 input features and one output variable i.e(Blood Sugar of a patient).
2. After a clear study and observation of the input features of the Diabetes Data. The BMI and the BP are the most co-related features for the prediction of the Blood sugar of the patient.

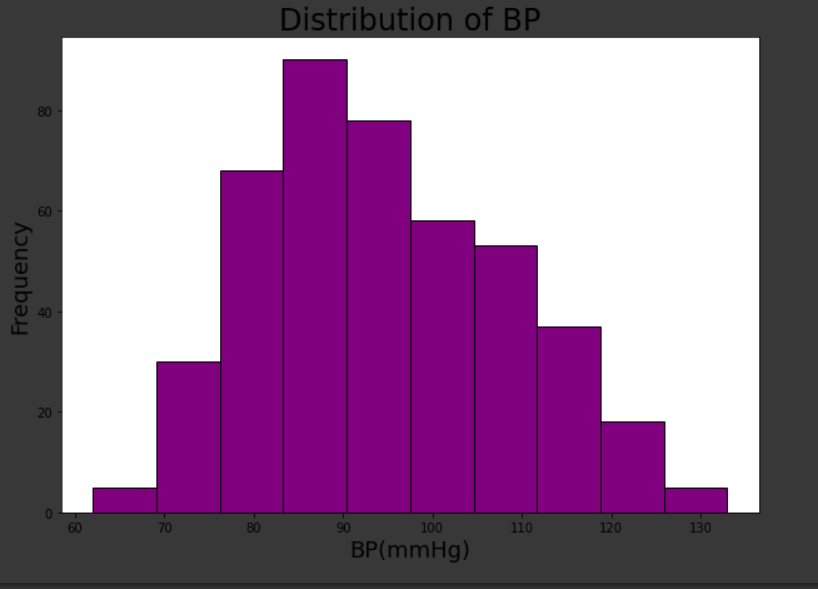
|  |  |
| --- | --- |
| BMI<18.5: | BELOW NORMAL WEIGHT |
| BMI>=18.5 AND <25: | NORMAL WEIGHT |
| BMI>=25 AND <30 | OVERWEIGHT |
| BMI>=30 AND <35 | CLASS 1 OBESITY |
| BMI>=35 AND <40 | CLASS 2 OBESITY |
| BMI>=40 | CLASS 3 OBESITY |

The people with **BMI >25** and also having **Hypertension** will have a high chance of getting diabetes.

Distribution of BMI data in the diabetes data set:



Distribution of BP data in the diabetes data set:



1. Moreover using the Pearson correlation coefficient the BMI and BP are having the highest values compared to other input features with the target output value.

In this project, we are going to use BMI and BP as two input features and Blood sugar becomes our output feature for training our model.

**Model Selection** :

As mention above this project’s aim is to the estimation of blood sugar levels using BMI and BP.

Since it is an estimation and dealing with two variables using a Multivariate linear regression is a better choice as a beginner.

#Here come’s the algorithm and code for the model

If we consider only BMI as our input feature using linear regression we achieved an accuracy of 98.3%.

If we consider BMI and BP as our input features using Multivariate linear regression we achieved an accuracy of 98.5%.

**Conclusion:**

|  |  |
| --- | --- |
| Model | Accuracy |
| Linear regression | 98.3% |
| Multivariate Linear regression | 98.45% |

Even though linear regression and Multivariate regression fetches a similar accuracy. Linear Regression is a better choice for the model selection because using less number of features we can achieve good accuracy.

How to run the code and use the application:

Two methods to run the code

1.Using Google Colab

2.Jupyter Notebook.

Here, I will tell you how to run the code in google colab

Step1: Open a google colab in your browser

Step2: Open a new notebook and click on the connect button which is located at the top right corner of the notebook.

Step3: Wait until it connects, make sure that you get RAM and disk space over there.

Step4: Copy the code and run it using CLTL+ENTER keys.