DIABETES PREDICTION LOGISTIC REGRESSION REPORT

MODEL REPORT

Analyst:
Rahul Verma
7973339701
rahulv23@iitk.ac.in

02

OBJECTIVE

My primary objective was ***to develop a robust predictive model to estimate the probability of diabetes "in females", in cheapest manner possible, without any lab access.*** This prediction will be based on a comprehensive analysis of diverse factors, including the number of pregnancies, blood pressure, skin thickness, body mass index (BMI), glucose levels, age, insulin levels, and the Diabetes Pedigree Function. By harnessing the power of these variables, our model seeks to uncover intricate patterns and relationships that can effectively differentiate between individuals with diabetes and those without. The overarching goal is to create a sophisticated and accurate tool that ***enhances female's ability to identify diabetes risk factors at home, without lab access.***

INDEPENDENT VARIABLES

- **Pregnancy** Frequency of pregnancy
- **Glucose** Concentration of plasma glucose (mg/dL)
- **BP** Diastolic blood pressure (mm Hg)
- **Skin** Tricep skinfold thickness (mm)
- **Insulin** Two-hour serum insulin (mu U/ml)
- **BMI** Body mass index (kg/m2)
- **Pedigree** A pedigree function for diabetes
- **Age** Age ((years))

DEPENDENT VARIABLES

- **Diabetes (Binary Variable)**

If a person is found diabetic, the value of variable = **1**, otherwise = **0**.

03

PROCESS FOLLOWED

- **EDA**
- **Data Cleaning**
- **Power Transform**
- **Machine Learning**
- **Logistic Regression**

Exploratory data analysis (EDA) : we used EDA to analyze and investigate data sets and summarize their main characteristics

SCATTER PLOTS

04 PROFILE REPORT

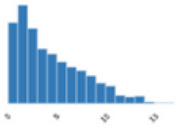
Pregnancies

Real number ($\mathbb{R}_{\geq 0}$)

HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION
ZEROS

Distinct	17
Distinct (%)	2.2%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Mean	3.845052083

Minimum	0
Maximum	17
Zeros	111
Zeros (%)	14.5%
Negative	0
Negative (%)	0.0%
Memory size	6.1 KiB

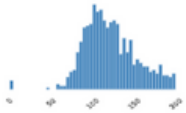


Glucose

Real number ($\mathbb{R}_{\geq 0}$)

Distinct	136
Distinct (%)	17.7%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Mean	120.8945312

Minimum	0
Maximum	199
Zeros	5
Zeros (%)	0.7%
Negative	0
Negative (%)	0.0%
Memory size	6.1 KiB



BloodPressure

Real number ($\mathbb{R}_{\geq 0}$)

HIGH CORRELATION
ZEROS

Distinct	47
Distinct (%)	6.1%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Mean	69.10546875

Minimum	0
Maximum	122
Zeros	35
Zeros (%)	4.6%
Negative	0
Negative (%)	0.0%
Memory size	6.1 KiB

SkinThickness

Real number ($\mathbb{R}_{\geq 0}$)

HIGH CORRELATION
ZEROS

Distinct	51
Distinct (%)	6.6%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Mean	20.53645833

Minimum	0
Maximum	99
Zeros	227
Zeros (%)	29.6%
Negative	0
Negative (%)	0.0%
Memory size	6.1 KiB

05 PROFILE REPORT

Insulin

Real number (ℝ≥0)

HIGH CORRELATION
ZEROS

Distinct	186
Distinct (%)	24.2%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Mean	79.79947917

Minimum	0
Maximum	846
Zeros	374
Zeros (%)	48.7%
Negative	0
Negative (%)	0.0%
Memory size	6.1 KiB

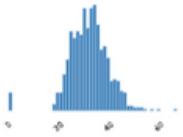
BMI

Real number (ℝ≥0)

HIGH CORRELATION
ZEROS

Distinct	248
Distinct (%)	32.3%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Mean	31.99257813

Minimum	0
Maximum	67.1
Zeros	11
Zeros (%)	1.4%
Negative	0
Negative (%)	0.0%
Memory size	6.1 KiB

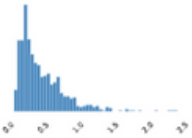


DiabetesPedig...

Real number (ℝ≥0)

Distinct	517
Distinct (%)	67.3%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Mean	0.4718763021

Minimum	0.078
Maximum	2.42
Zeros	0
Zeros (%)	0.0%
Negative	0
Negative (%)	0.0%
Memory size	6.1 KiB



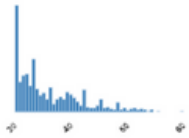
Age

Real number (ℝ≥0)

HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION

Distinct	52
Distinct (%)	6.8%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Mean	33.24088542

Minimum	21
Maximum	81
Zeros	0
Zeros (%)	0.0%
Negative	0
Negative (%)	0.0%
Memory size	6.1 KiB



06

DATA DESCRIPTION AND SOURCE

We use the Pima Indian dataset, made available by the National Institute of Diabetes at the Johns Hopkins University, as a test case to predict the risk factors associated with diabetes. The Pima are American Indians that live along the Gila River and Salt River in Southern Arizona. This dataset consists of 768 rows with 9 features.

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1
...	...	...	...	...	...	...	...	...	...
763	10	101	76	48	180	32.9	0.171	63	0
764	2	122	70	27	0	36.8	0.340	27	0
765	5	121	72	23	112	26.2	0.245	30	0
766	1	126	60	0	0	30.1	0.349	47	1
767	1	93	70	31	0	30.4	0.315	23	0

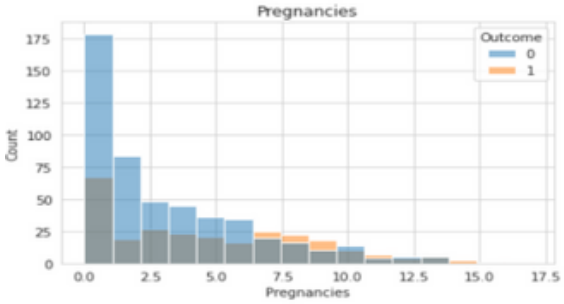
POWER TRANSFORM AND STANDARD SCALING

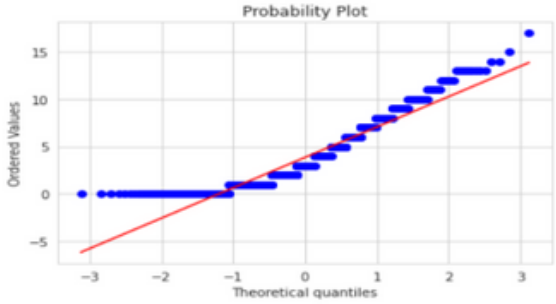
From the EDA we can observe the range for different variables is very different. For e.g. Insulin can take a value up to 846, where as maximum value for variables like DiabetesPedigreeFunction in the dataset is 2.42. If we use the values as given, they will affect the beta values and might mislead the observer in comparing the effect due to a certain independent variable on the dependent variable as compared to effect due to other independent variable. To avoid this we scaled the data to a standard normal distribution.

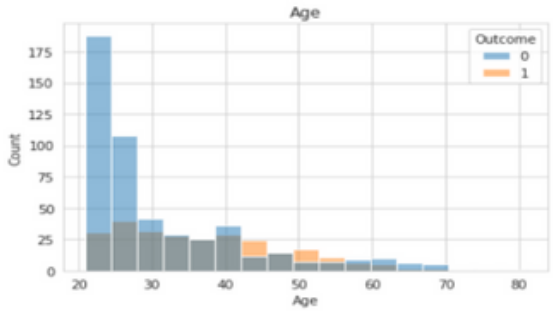
Another observation from the EDA is that the data for some independent variables is skewed. To handle this we used the Power Transform.

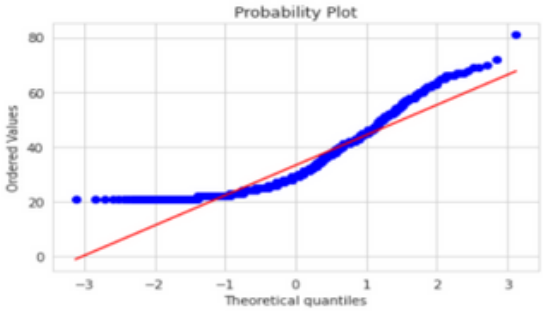
07 POWER TRANSFORM AND STANDARD SCALING

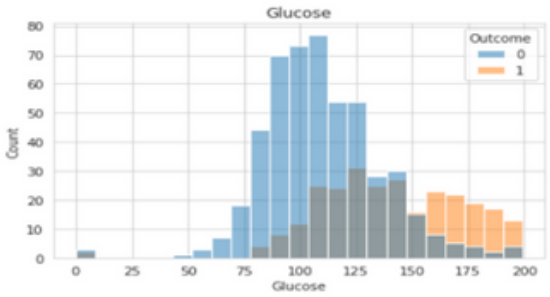
DATA BEFORE SCALING AND POWER TRANSFORM

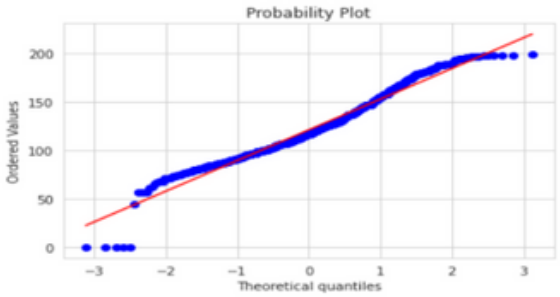


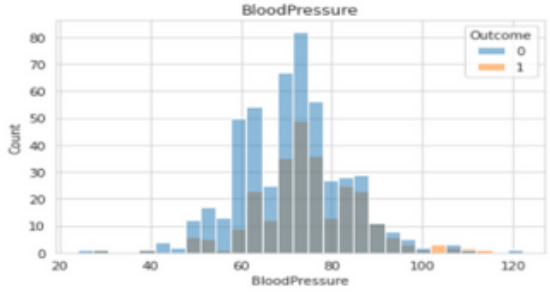


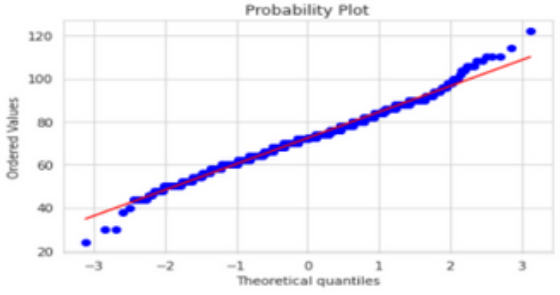


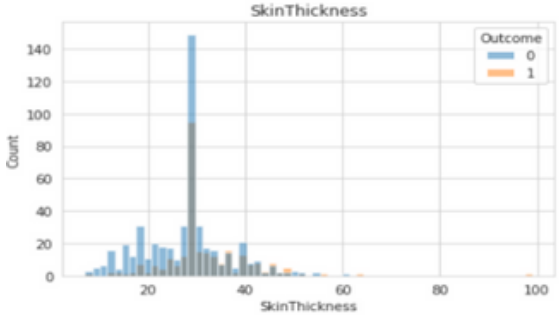


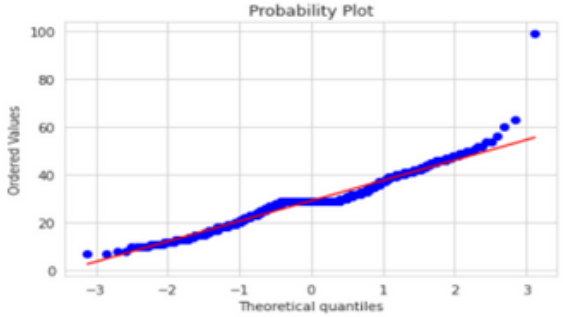




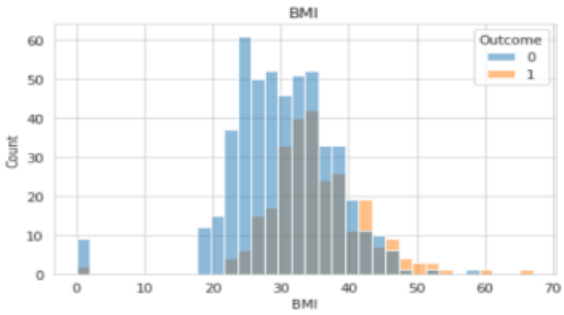




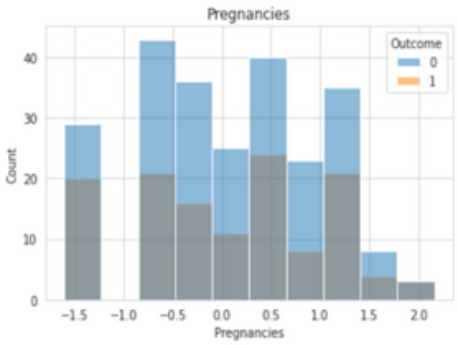


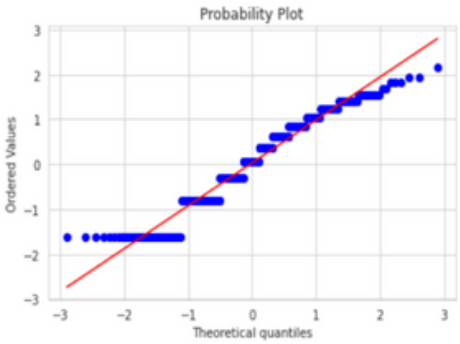


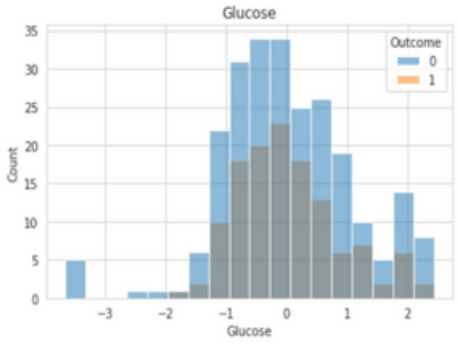
08 POWER TRANSFORM AND STANDARD SCALING

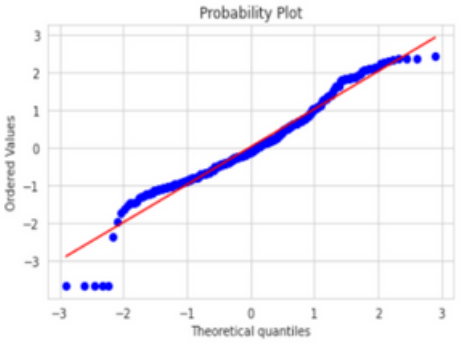


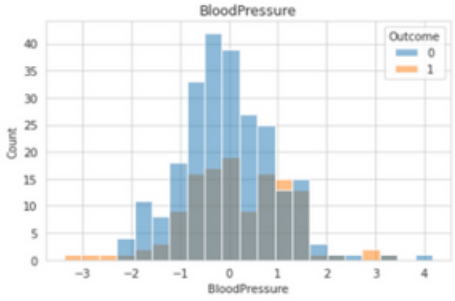
DATA AFTER SCALING AND POWER TRANSFORM

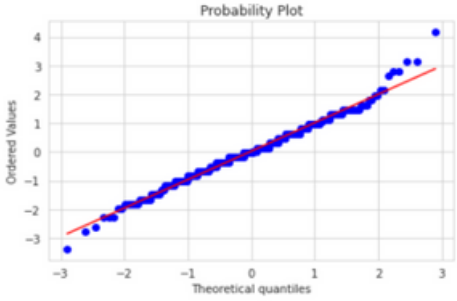




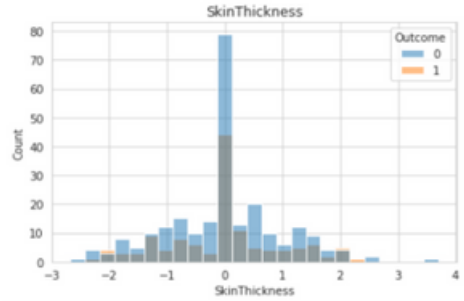


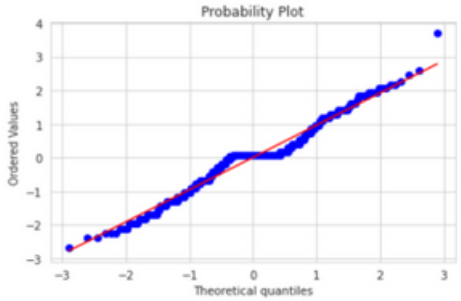


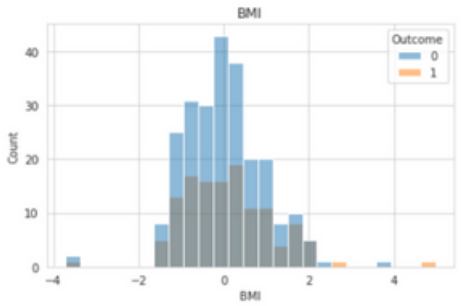


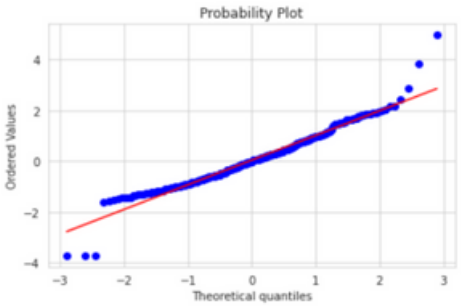


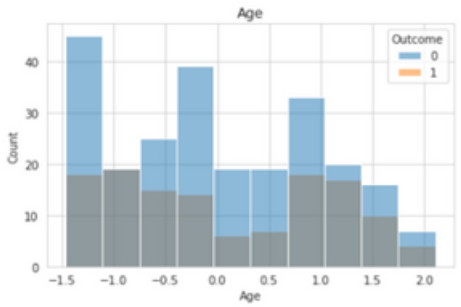
09 POWER TRANSFORM AND STANDARD SCALING

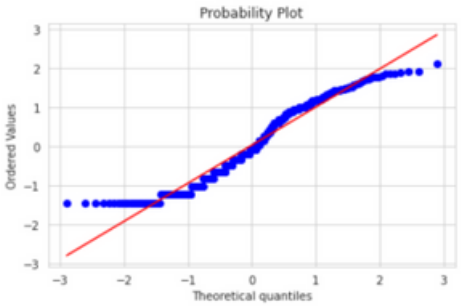








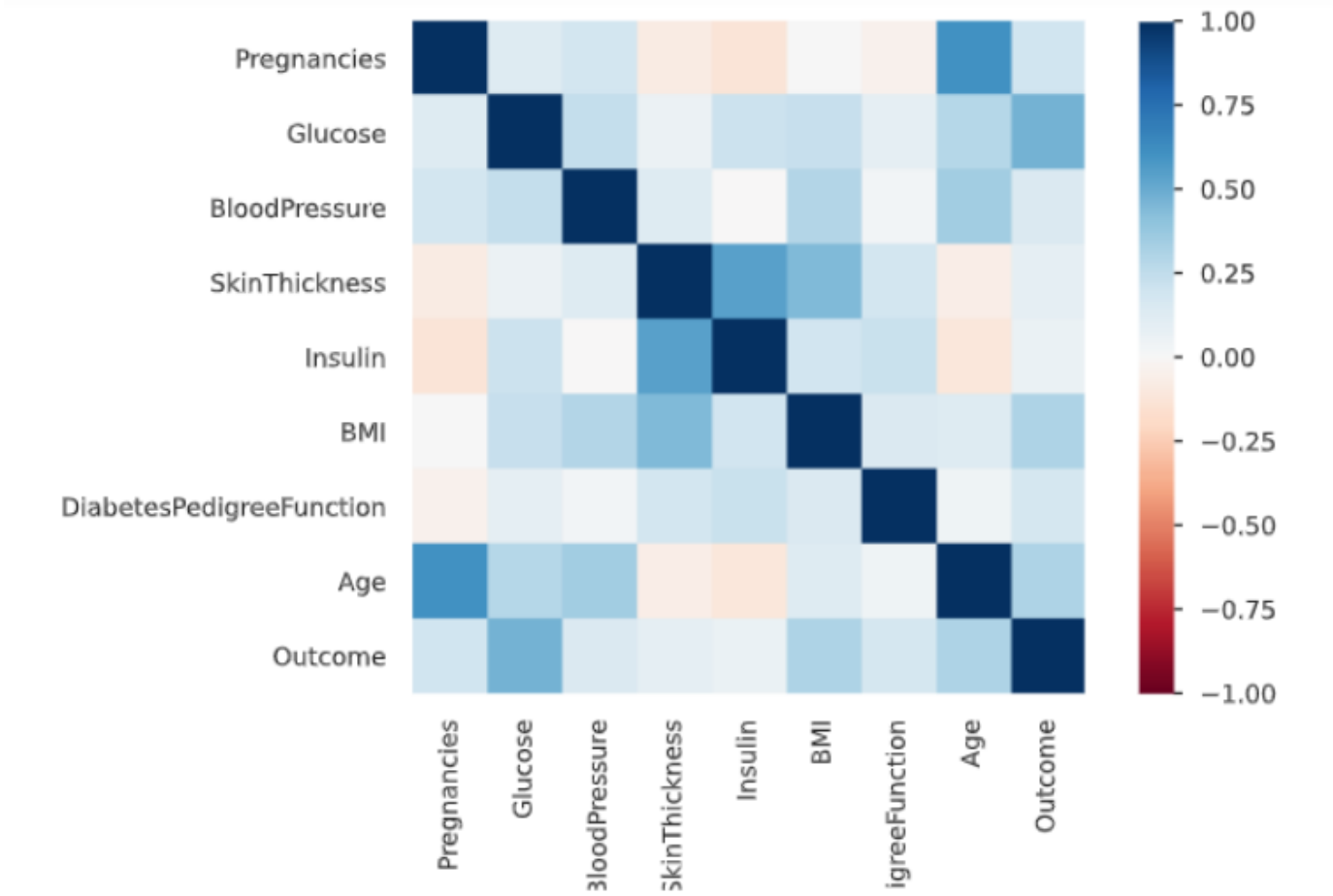




10

CORRELATION MATRIX

THE FOLLOWING CORRELATION MATRIX WAS OBTAINED:

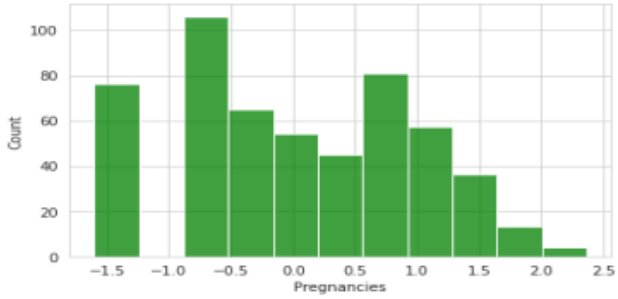


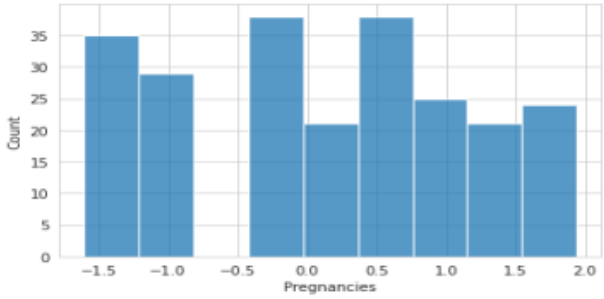
IT CAN BE OBSERVED FROM THE ABOVE MATRIX THAT THE CORRELATION BETWEEN THE INSULIN AND SKIN THICKNESS IS HIGH AND THEY BOTH ARE CORRELATED WITH THE DEPENDENT VARIABLE. SO TO AVOID MULTICOLLINEARITY WE SHOULD DROP ONE OF THESE VARIABLES. WE DECIDED TO DROP INSULIN BECAUSE IT IS EASIER AND CHEAPER TO MEASURE SKIN THICKNESS WHEREAS MEASURING INSULIN REQUIRES THE LAB ACCESS.

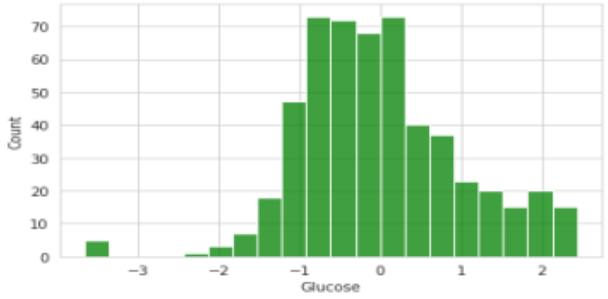
11 TRAINING AND TESTING OF THE MODEL

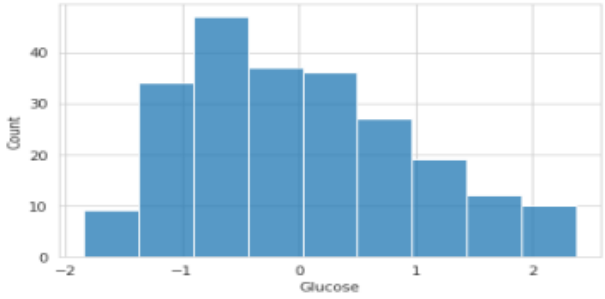
WE TRAIN THE MODEL ON 70% OF THE DATA AND TEST ON THE REMAINING 30% TO CHECK THE ACCURACY OF THE MODEL.

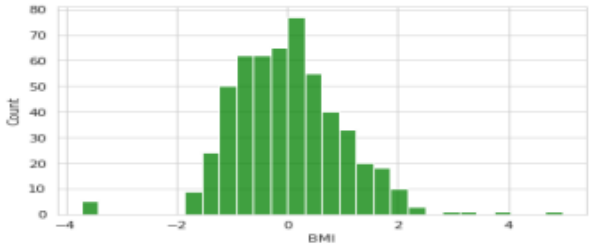
IN GREEN WE HAVE OUR TRAINING DATA AND THE TEST DATA IS SHOWN IN BLUE.

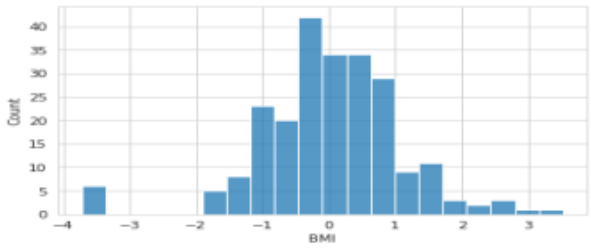


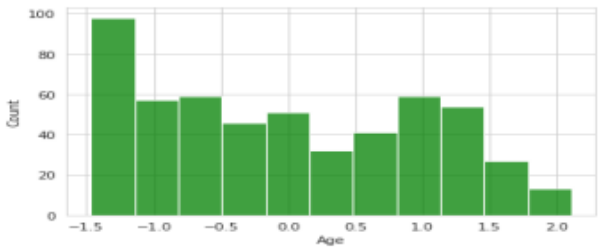


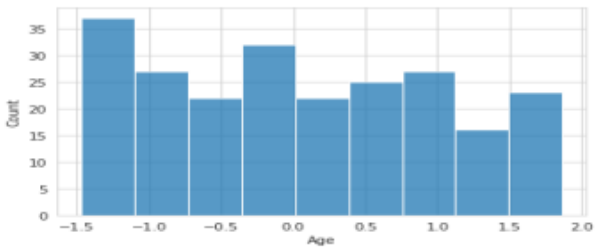


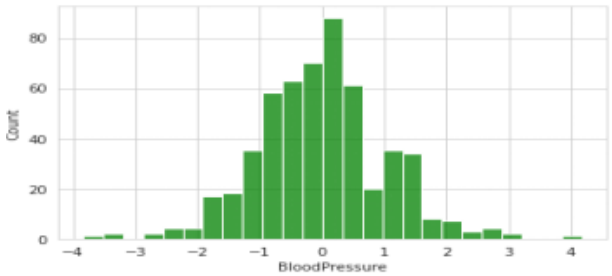


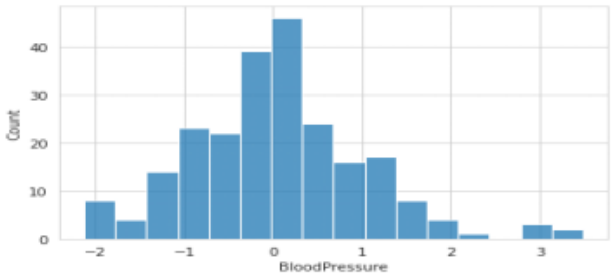








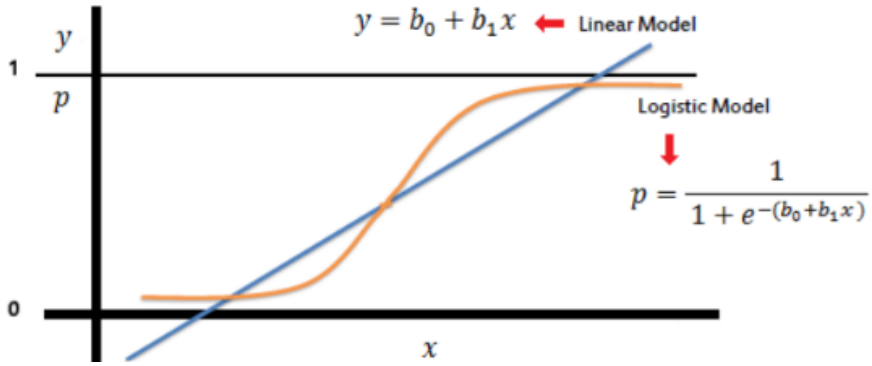




12

LOGISTIC REGRESSION

WE DID LOGISTIC REGRESSION USING LOGIT FUNCTION **F(X)=1/(1+EXP(-(B2+B1X))**. WE WILL PREDICT THAT THE PERSON IS NOT DIABETIC IF PROBABILITY IS LESS THAN 0.5 AND PERSON IS DIABETIC IF PROBABILITY IS MORE COMES OUT TO BE MORE THAN 0.5.



MODEL 1

Variable	Beta Value	P-Value
Const	-8.0777	0.000
AGE	0.0386	0.009
BMI	0.1145	0.000
Skin Thickness	-0.0052	0.200
Pregnancies	0.0682	0.068

INITIALLY WE HAD 7 VARIABLES. OUR MAIN OBJECTIVE TO MAKE THIS MODEL WAS THAT , THE PERSON CAN CHECK WHETHER SHE IS DIABETIC OR NOT WITHOUT ANY LAB ACCESS AND AT LESS COST. SO, IN OUR FIRST MODEL WE DROPPED 2 VARIABLE WHICH WERE **INSULIN** AND **DIABETES PEDIGREE FUNCTION** BECAUSE BOTH OF THEM REQUIRE LAB ACCESS.
ALL THE VARIABLES CAME OUT TO BE SIGNIFICANT EXCEPT **PREGNANCIES**.

13

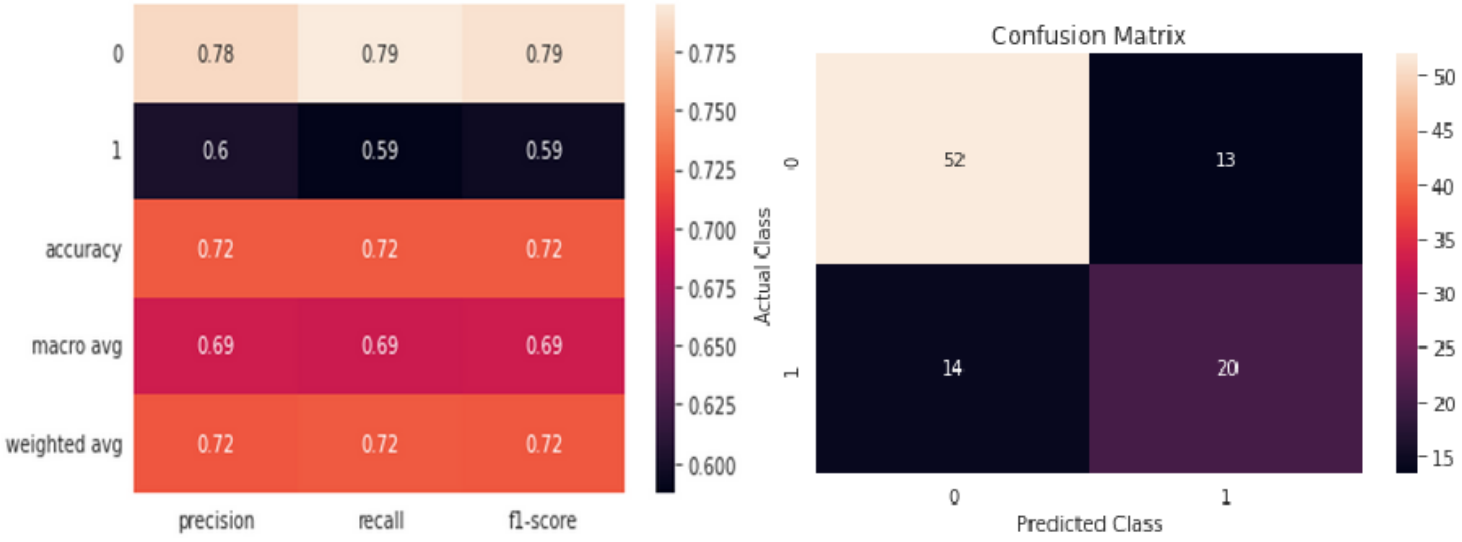
LOGISTIC REGRESSION

MODEL 2

Variable	Beta Value	P-Value
Const	-9.7077	0.000
AGE	0.0313	0.003
BMI	0.1019	0.000
Skin Thickness	-0.0011	0.000
Glucose	0.0363	0.000
Pregnancies	0.0656	0.028

IN OUR MODEL 2, WE INCLUDE ANOTHER VARIABLE WHICH IS GLUCOSE. INCLUDING GLUCOSE AS OUR INDEPENDENT MODEL, THE VARIABLE PREGNANCIES ALSO BECAME SIGNIFICANT. ALSO OUR PSEUDO R2 INCREASED FROM **0.214** TO **0.276** AS COMPARED TO MODEL 1.

THEN, WE MADE CONFUSION MATRIX. OUR ACCURACY CAME OUT TO BE **0.72**.



CONCLUSION:

WE CONCLUDED THAT OUR **MODEL 2 WAS BETTER MODEL AS COMPARED TO MODEL 1**. ALSO, AS THIS MODEL IS FOR PREDICTING WETHER THE PERSON IS DIABETIC OR NOT, SO WE SHOULD FOCUS ON REDUCING THE PREDICTION WHICH ARE FALSE NEGATIVE OR WE CAN SAY THAT REDUCING TYPE 2 ERROR WILL IMPROVE OUR MODEL.

14

FINAL SUMMARY AND OUTCOME

1. **ACCURACY AND RELIABILITY:**
THIS MODEL EXHIBITS HIGH ACCURACY, AND ITS EFFORTS TO MINIMIZE TYPE 2 ERRORS, SPECIFICALLY REDUCING THE LIKELIHOOD OF MISCLASSIFYING A DIABETIC FEMALE AS NON-DIABETIC, ENHANCE ITS OVERALL RELIABILITY.

2. **COST-EFFECTIVENESS AND ACCESSIBILITY:**
MOREOVER, THE MODEL'S AFFORDABILITY AND MINIMAL COST MAKE IT ACCESSIBLE FOR FEMALES SEEKING TO ASSESS THEIR RISK OF DIABETES, RENDERING IT A VALUABLE TOOL FOR INTEGRATION INTO VARIOUS SOFTWARE OR WEBSITES FOCUSED ON THIS HEALTH TOPIC.

3. **VARIABLES AND COST CONSIDERATIONS:**
HOWEVER, CONCERNS MAY ARISE REGARDING THE PRACTICALITY OF THE APPROACH WHEN CONSIDERING HOW USERS CAN OBTAIN THEIR BMI, SKIN THICKNESS, AND GLUCOSE LEVEL AT A LOW COST, WHEREAS VARIABLES SUCH AS AGE AND PREGNANCIES INCUR LITTLE TO NO EXPENSE.
REPY TO THAT IS,

- **CALCULATING BODY MASS INDEX (BMI)** REQUIRES A STRAIGHTFORWARD AND COST-FREE FORMULA: BMI = WEIGHT IN KILOGRAMS / (HEIGHT IN METERS)^2.
- **SKIN THICKNESS** ASSESSMENT CAN BE CONDUCTED INEXPENSIVELY USING READILY AVAILABLE SKINFOLD CALIPERS. ONCE MEASURED, SKIN THICKNESS DATA REMAINS RELATIVELY STABLE OVER TIME.
- ADDITIONALLY, **GLUCOSE LEVELS** CAN BE DETERMINED AFFORDABLY THROUGH THE USE OF BLOOD GLUCOSE METERS, WHICH ARE WIDELY ACCESSIBLE AT MOST PHARMACIES.

4. **IMPROVING MODEL RELIABILITY AND REGIONAL SPECIFICITY:**
FURTHERMORE, THE CONTINUOUS AUGMENTATION OF DATA OVER TIME CONTRIBUTES TO THE ONGOING ENHANCEMENT OF THIS MODEL'S RELIABILITY, ENSURING A SUSTAINED IMPROVEMENT IN ACCURACY. ADDITIONALLY, THE INCORPORATION OF A NEW FEATURE RELATED TO REGIONAL FOOD HABITS COULD TAILOR THE MODEL TO SPECIFIC GEOGRAPHIC AREAS. FOR INSTANCE, JAPANESE CUISINE, KNOWN FOR ITS MINIMAL USE OF SUGAR IN TRADITIONAL DISHES AND EMPHASIS ON UMAMI AND SAVORY FLAVORS, CONTRASTS WITH THE MORE SUGAR-ORIENTED NATURE OF INDIAN CUISINE. THE INTRODUCTION OF REGIONAL SPECIFICITY WOULD ADD A NUANCED LAYER TO THE MODEL'S PREDICTIVE CAPABILITIES.