**DATA ANALYTICS GROUP PROJECT -**

**Predicting through diagnostic measurement whether a patient has diabetes**

# **1.Introduction**

Diabetes is a low blood sugar level occurring in a person with diabetes mellitus. According to the National Electronic Injury Surveillance System -All Injury Program (NEISS-AIP) and based on sample examined between 2004 and 2005, an estimated 55,819 cases involved presence of high blood sugar level. Diabetic Hypoglycemia occurs when a treatment to reduce the high blood sugar of diabetes inaccurately matches the body’s physiological need, and therefore causes the glucose to fall to a below normal level. Diabetes can be infected in a person due to couple of reasons like high glucose level, hereditary factor, stress and many more. In our scenario, one of the primary variable is blood sugar level. Blood sugar level is the deciding factor that says whether a person has diabetes or no. Other variables which can also be taken into consideration are Age, High Blood Pressure, Weight, Family History, Abnormal Cholesterol level and so on.

# **Motivation**

1. To Predict if a patient has diabetes based on the diagnostic measurements specified.

2. To Identify the most important factors that contribute to Diabetes.

3. To Analyze if the Age of the person has any bearing on increasing the risk of Diabetes.

4. To Analyze if the Diabetic Pregnant women and their babies have high risk of diabetes when compared to Non- Diabetic Pregnant women.

5. To Analyze if the patients having diabetic relatives are more vulnerable to acquire Diabetes when compared to other patients having non - diabetic relatives.

# **2.Data Description**

We decide to use the diabetes dataset from **PIMA INDIANS DIABETES DATASET, KAGGLE.**

**Total Number of Dataset: 1268.**

Our dataset was divided into 2 types of data:

**Numeric**

* Pregnancies.
* Glucose.
* Blood Pressure.
* Insulin.
* Skin Thickness.
* BMI.
* Age.
* Diabetes Pedigree Function.

**Categorical**

* Outcome.

# **3.Research Problems and Solutions**

After the dataset was collected the first problem was to get actual data out of raw data and second was to decide which model will be best fitted for our dataset. To get the actual data, we performed data cleaning and our final dataset had 1268 rows. For best fitted model, we performed model evaluation. Below we have elaborated as to how we did that.

# **4.Model Learning**

# 4.1 Data Processing

**Collection of Data:**

* Using KAGGLE we collected diabetes dataset from PIMA INDIANS.
* After performing data cleaning the data which we had with us counted to 1268.

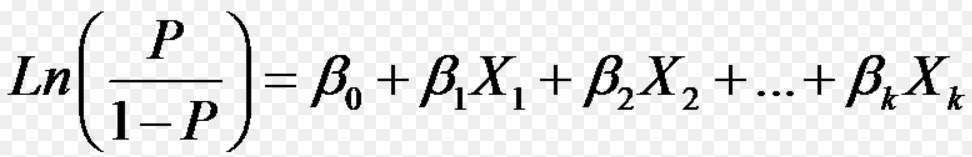
**Model Evaluation:**

* It is a fundamental piece of the model advancement prepare.
* Assessing model execution with the information utilized for preparing is not adequate in information mining since it can without much of a stretch create overoptimistic and overfitted models.
* We tried testing our model with Forward Model, Backward Model and Stepwise Regression.
* **Forward modelling** which includes beginning without any factors in the model, testing the option of every variable utilizing a picked show fit rule, including the variable (assuming any) whose consideration gives the most factually huge change of the fit, and rehashing this procedure until none enhances the model to a measurably critical degree.
* **Backward modelling** which includes beginning with all competitor factors, testing the erasure of every variable utilizing a picked show fit foundation, erasing the variable (assuming any) whose misfortune gives the most measurably unimportant decay of the model fit, and rehashing this procedure until no further factors can be erased without a factually huge loss of fit.
* **Stepwise regression** is a strategy for fitting relapse models in which the decision of prescient factors is completed by a programmed system.
* Performing evaluation steps with the present dataset we concluded that **Stepwise Regression** is best fitted for our model.

# 4.2 Data Analytics Tasks and Processes

**Logistic regression**, or logit regression, or logit model  is a [regression](https://en.wikipedia.org/wiki/Regression_analysis) model where the [dependent variable](https://en.wikipedia.org/wiki/Dependent_and_independent_variables)  is [categorical](https://en.wikipedia.org/wiki/Categorical_variable). Logistic regression generally known as Generalized linear model is a modelling technique used for classification problem. Logistic regression measures the relationship between the categorical dependent variable and one or more independent variables by estimating probabilities using a [logistic function](https://en.wikipedia.org/wiki/Logistic_function), which is the cumulative [logistic distribution](https://en.wikipedia.org/wiki/Logistic_distribution). Logistic regression allows prediction of discrete variables by a mix of continuous and discrete predictors. Logistic Regression addresses the same questions that discriminant function analysis and multiple regression do but with no distributional assumptions on predictors. Logistic Regression measures relation between categorical dependent variable and one or more independent variable. It estimates probabilities using logistic function.

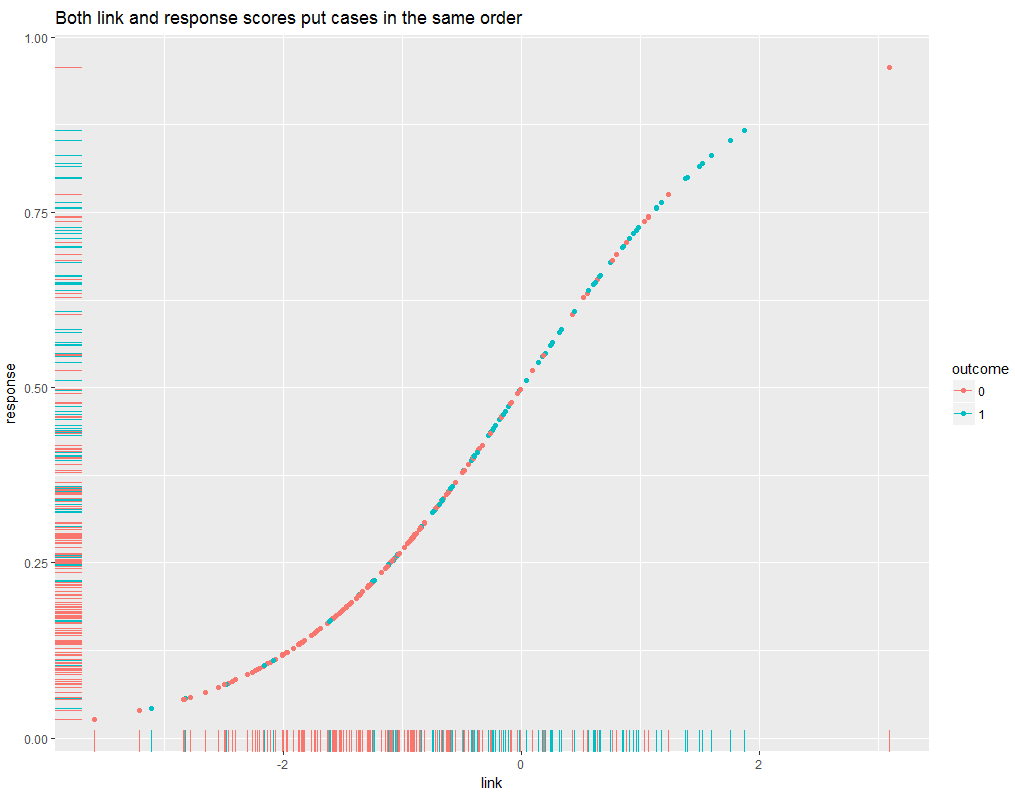
Logistic regression generates the coefficients, standard errors and significance levels of a formula to predict a logit transformation of the probability of presence of the characteristic of interest:



Logistic regression goal is to find the best fitting model. logistic regression model describes the relationship between the dependent variable and a set of independent variables in a defined manner. Logistic Regression is mainly used when the dependent variable is found to be “limited.” With yes/No types or Binary range of values. Binary logistic regression is a type of regression analysis where dependent variable is a dummy variable is either 0 or 1.

In our project we chose logistic regression as the output variable ( dependent variable ) ‘Output’ has Binary range of values either 0 or 1. We found the relationship between dependent variable ‘Output’ which is Categorical and Independent variables ‘Pregnancies’, ‘Glucose’, ‘Blood Pressure’ , ‘Skin Thickness’ , ‘Insulin’ , ‘BMI’ , ‘Diabetes Pedigree Function’ and ‘Age’ which are Numeric.

The logistic regression model chooses parameters that maximize the likelihood of observing the sample values rather than choosing parameters that minimize the sum of squared errors estimation as utilized in linear model. The logistic regression generalizes linear regression by allowing the linear model to be related to the response variable via a link function. The link function provides the relationship between the linear predictor and the mean of the distribution function. It makes prediction using probability meaning the values are from 0 to 1. We can assume that if the predicted value is more than 0.5, then it can be classified as 1. If the value is less than 0.5, then it can be classified as 0.



**Figure 1: LOGISTIC REGRESSION GRAPH**

# **5.Evaluations and Results**

# 5.1 Evaluation Methods

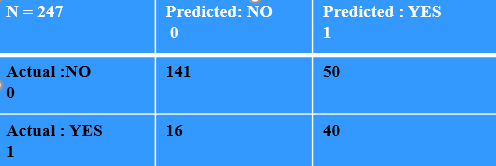
**CONFUSION MATRIX**

Confusion Matrix is a table that is regularly used to portray the execution of a grouping model (or "classifier") on an arrangement of test information for which the genuine qualities are known.

We Predicted our Logistic Regression Model Accuracy using **Confusion Matrix**.

With N values to be 247 observations, and found:

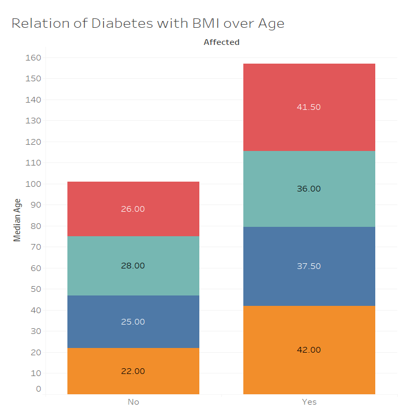
* True Positives values to be **141**
* False Negative values to be **50**
* False Positive values to be **16**
* True Negative values to be **40**

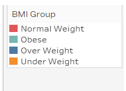
****

# 5.2 Results and Findings

**RELATION OF DIABETES WITH BMI OVER AGE**

This graph depicts the relationship of the BMI values with Median of Age Values.

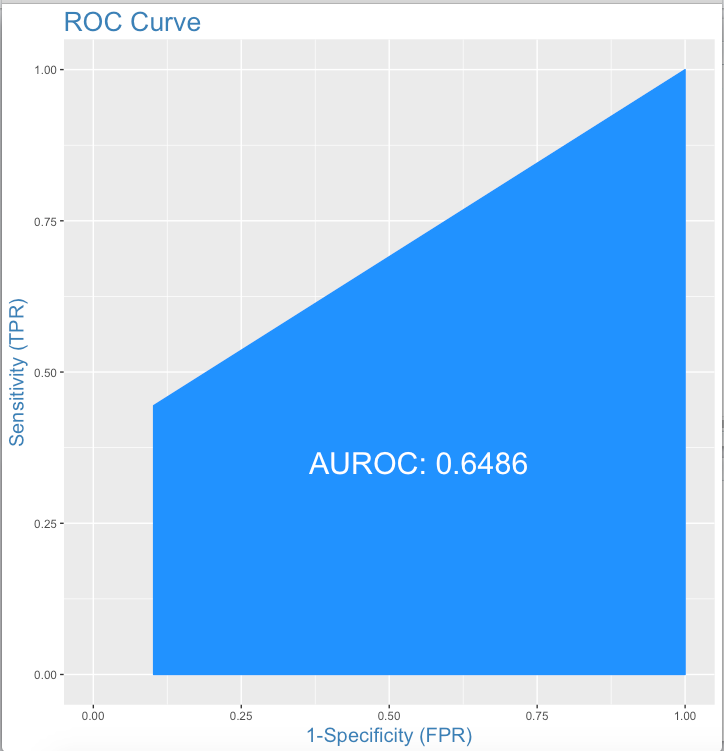




**Figure 2: BMI VS AGE**

**ROC CURVE**

ROC curve analysis is performed to predict the power of the model's predicted values which discriminates the range between positive and negative cases. ROC curve is generally quantified by the Area under the ROC curve The AUC is also referred to as concordance index, it has values that varies from 0.5 which is characterized as discriminating power not better than chance to 1.0 which is characterized as perfect discriminating power. Traces the percentage of the True Positives accurately from the given logit model.ROC is considered most ordinarily utilized approach to picture the execution of a paired classifier, and AUC is (seemingly) the most ideal approach to compress its execution in a solitary number. It traces the percentage of the True Positives accurately from the given logit model.

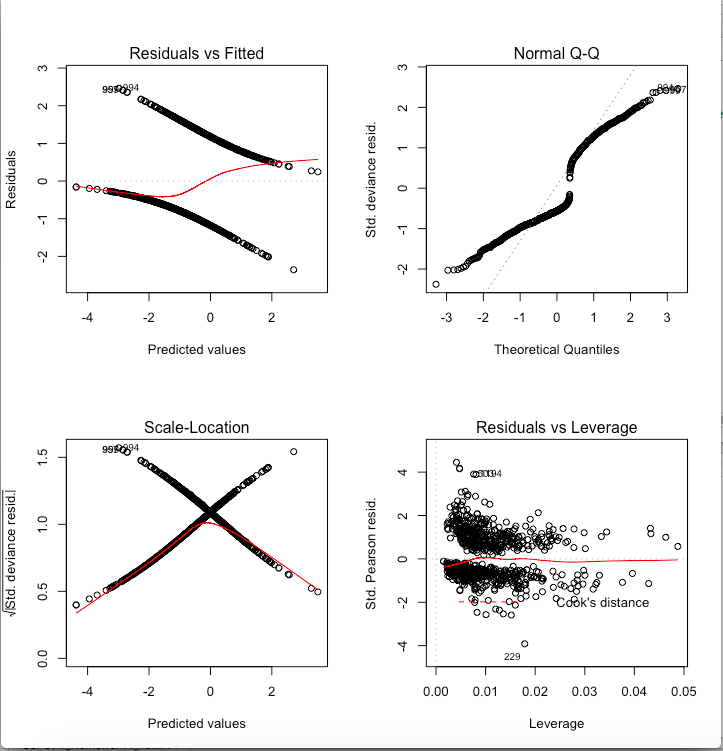


**Figure 3: SENSITIVITY VS SPECIFICITY**

**RESIDUAL ANALYSIS**

It is a chart that demonstrates the residuals on the vertical hub and the autonomous variable on the level hub. If the focuses in a leftover plot are haphazardly scattered around the level pivot, a direct relapse model is fitting for the information; generally, a non-straight model is more proper.

Underneath the table on the left shows sources of info and yields from a straightforward direct relapse examination, and the graph on the privilege shows the leftover (e) and free factor (X) as a remaining plot.



**Figure 4: RESIDUAL PLOTS**

**CORRELATION PLOTS**

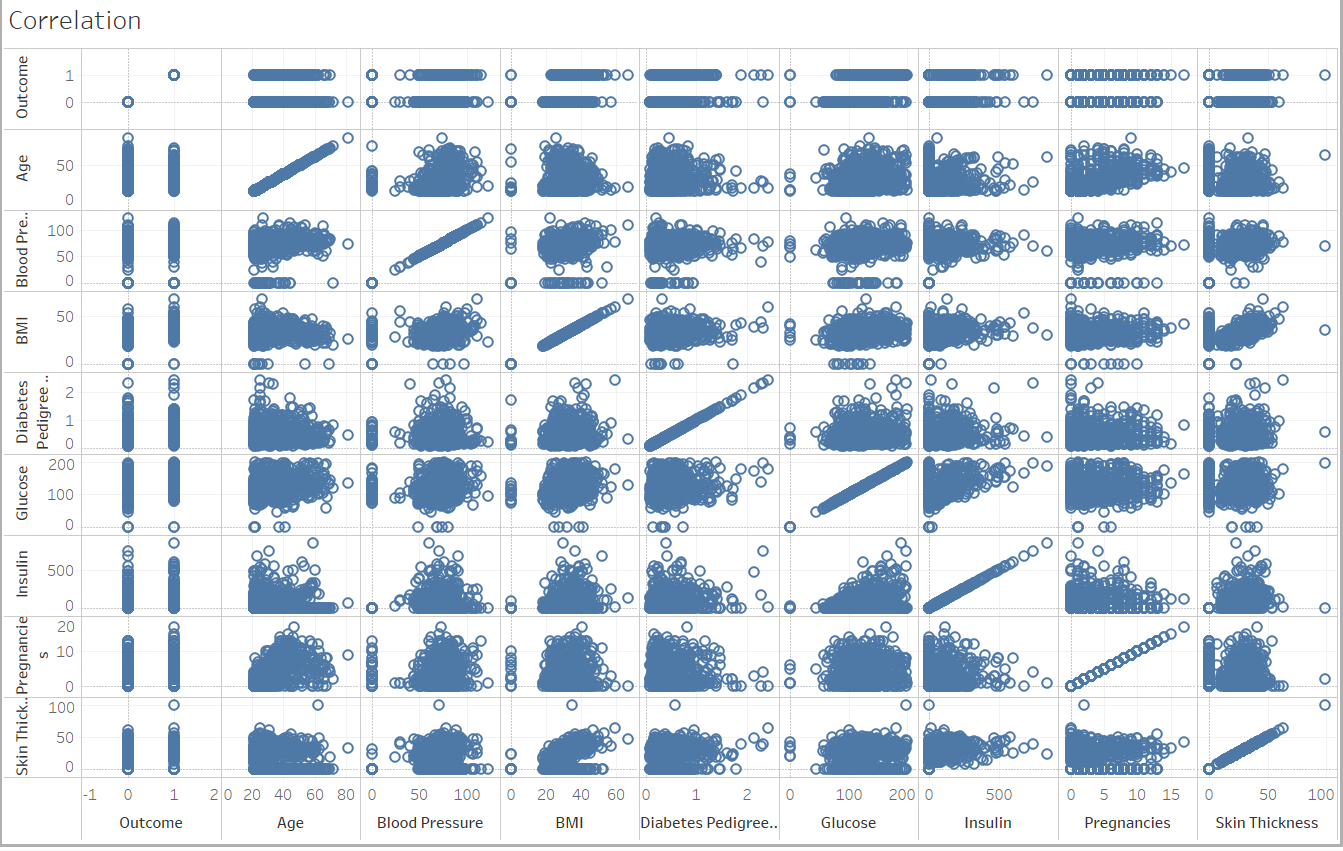
All correlations have two properties: quality and heading. The quality of a connection is dictated by its numerical esteem. The bearing of the correlation is dictated by whether the connection is positive or negative.

**Positive correlation**: Both factors move in a similar course. At the end of the day, as one variable builds, the other variable likewise increments. As one variable declines, the other variable additionally diminishes.

i.e., years of training and yearly compensation are emphatically associated.

**Negative correlation**: The factors move in inverse headings. As one variable expands, the other variable abatements. As one variable declines, the other variable increments.

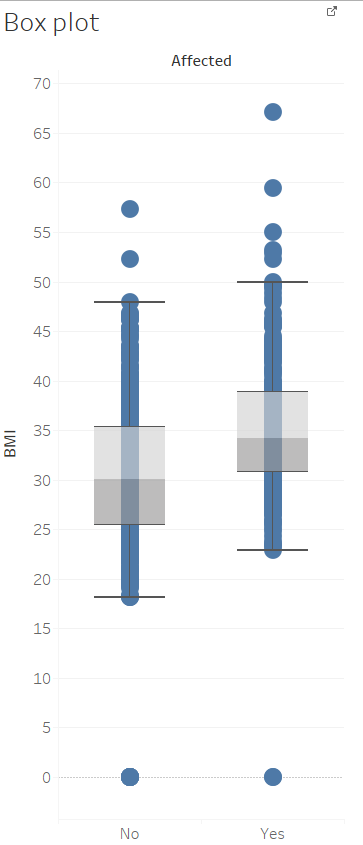
i.e., hours spent dozing and hours spent alert are contrarily related.



**Figure 5: DIFFERENT FACTORS CAUSING DIABETES**

**BOX PLOT**

The middle (center quartile) denote the mid-purpose of the information and is appeared by the line that partitions the crate into two sections. A large portion of the scores are more prominent than or equivalent to this esteem and half are less. The center "box" speaks to the center half of scores for the gathering.



**Figure 6: BMI VS AFFECTED**

# **6.Conclusins and Future Work**

* By prediction, accuracy rate of diabetes was found to be 73.279.
* Based on Analysis, the influential diabetic factors are concluded to be glucose, BMI, Pregnancies, Blood Pressure and Diabetes Pedigree Function.
* Based on our fitted model, Age is not a parameter to determine the risk of diabetes.
* The best fitted model was concluded to be stepwise regression model.